Sensorimotor Learning and Simulation of Experience as a Basis for the Development of Cognition in Robotics

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)
im Promotionsfach Informatik

Mathematisch-Naturwissenschaftliche Fakultät II
der Humboldt-Universität zu Berlin

von M.Sc. **Guido Schillaci**

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II:
Prof. Dr. Elmar Kulke

Gutachter(innen):
1. Prof. Dr. Verena V. Hafner (Humboldt-Universität zu Berlin)
2. Prof. Dr. Bruno Lara (Universidad Autónoma del Estado de Morelos, Mexico)
3. Prof. Dr. Angelo Cangelosi (University of Plymouth, UK)

Eingereicht an der 01. November 2013
Tag der Verteidigung: 16. Dezember 2013

A mia figlia Alice
per la sua nascita

Abstract

State-of-the-art robots are still not properly able to learn from, adapt to, react to unexpected circumstances, and to autonomously and safely operate in uncertain environments. Researchers in developmental robotics address these issues by building artificial systems capable of acquiring motor and cognitive capabilities by interacting with their environment, inspired by human development. This thesis adopts a similar approach in finding some of those basic behavioural components that may allow for the autonomous development of sensorimotor and social skills in robots.

Here, sensorimotor interactions are investigated as a mean for the acquisition of experience. Experiments on exploration behaviours for the acquisition of arm movements, tool-use and interactive capabilities are presented. The development of social skills is also addressed, in particular of joint attention, the capability to share the focus of attention between individuals. Two prerequisites of joint attention are investigated: imperative pointing gestures and visual saliency detection.

The established framework of the internal models is adopted for coding sensorimotor experience in robots. In particular, inverse and forward models are trained with different configurations of low-level sensory and motor data generated by the robot through exploration behaviours, or observed by human demonstrator, or acquired through kinaesthetic teaching. The internal models framework allows the generation of simulations of sensorimotor cycles. This thesis investigates also how basic cognitive skills can be implemented in a humanoid robot by allowing it to recreate the perceptual and motor experience gathered in past interactions with the external world. In particular, simulation processes are used as a basis for implementing cognitive skills such as action selection, tool-use, behaviour recognition and self-other distinction.

Zusammenfassung

Heutige Roboter sind nur begrenzt in der Lage etwas zu erlernen, sich unerwarteten Umständen anzupassen oder auf diese zu reagieren. Als Antwort auf diese Fragen, develomental robotics setzt sich den Aufbau eines künstlichen Systems zum Ziel, das motorische und kognitive Fähigkeiten analog zur menschlichen Entwicklung durch Interaktion mit der Umgebung entwickeln kann.

In dieser Arbeit wird ein ähnlich Ansatz verwendet, mit Hilfe dessen grundlegende Verhaltenskomponenten identifiziert werden sollen, die eine autonome Aneignung motorischer und kognitive Fähigkeiten durch die Roboter ermöglichen knnten. Diese Arbeit untersucht die sensomotorische Interaktion als Mittel zur Schaffung von Erfahrungen. Es werden Experimente zu explorative Verhaltensweisen zur Aneigung von Arbewegungen, der Werkzeugnutzung und von interaktiven Fähigkeiten vorgestellt. In diesem Rahmen wird auch die Entwicklung sozialer Fähigkeiten, insbesondere durch joint attention, behandelt. Dabei werden zwei Vorraussetzugen zu joint attention untersucht: Zeigegesten und Erkennung von visueller Salienz.

Dabei wurde das Framework der interen Modelle für die Darstellung von sensomotorischen Erfahrungen angewendet. Insbesondere wurden inverse und Vorwärtsmodelle mit unterschiedlichen Konfigurationen am sensorischen und motorischen Daten, die vom Roboter durch exploratives Verhalten, durch Beobachtung menschliche Vorführern, oder durch kinästhetisches Lehren erzeugt wurden geschult. Die Entscheidung zu Gunsten dieses Framework wurde getroffen, da es in der Lage ist, sensomotorische Zyklen zu simulieren. Diese Arbeit untersucht, wie grundlegende kognitive Fähigkeiten in einen humanoiden Roboter unter Berücksichtigung sensorischer und motorischer Erfahrungen implementiert werden können. Insbesondere wurden interne Simulationsprozesse für die Implementierung von Kognitivenfahigkeiten wie die Aktionsauswahl, die Werkzeugnutzung, die Verhaltenserkennung und die Self-Other distinction, eingesetzt.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Prof. Verena V. Hafner, for her confidence in my research and her support during my Ph.D. studies. I am honoured I had the opportunity to join the EU-FP7 Marie Curie ITN INTRO network and to work with all the researchers that took part in INTRO's activities. I would like to thank Prof. Thomas Hellström, coordinator of INTRO, for his efforts in making INTRO a great experience and for the help and nice hospitality at the UmeåUniversity (Sweden), where I spent almost 3 months for my secondment. I am very grateful to all the professors and scientists that made INTRO's activities very stimulating and constructive. Particular thanks go to the INTRO's Early Stage Researchers and Experienced Researchers for making our frequent meetings and trips an unforgettable experience: Saša Bodiroža, Guillaume Doisy, Maria Elena Giannaccini, Bo Li, Aleksander Jevtic, Ahmad Mosallam, Benjamin Fonooni, Roy Someshwar, Alex Koslov. I am looking forward to the next meet-up!

I am grateful to the Humboldt-Universität zu Berlin who hosted my Ph.D. studies and to the members of the Kognitive Robotik Group, for their feedbacks and for the nice times in the lab, in particular: Siham Al-Rikabi, Ferry Bachmann, Oswald Berthold, Christian Blum, Damien Drix, Ivana Kajič, Philipp Rhan and to the Berlin United - Nao Team Humboldt people, in particular to Marcus Scheunemann, Heinrich Mellmann, Claas Ritter and Thomas Krause. A big *thank you* goes to Renate Zirkelbach, for helping me, with great patience, in dealing with the HU bureaucracy and for the funny moments we had while carrying those tasks.

Special thanks go also to my second supervisor, Prof. Bruno Lara (Universidad Autònoma del Estado de Morelos, Mexico), for his support and his advices in the development of the ideas and experiments presented in this thesis and for the funny time we had together out of the lab. I am also thankful to Marc Grosjean (Leibniz Research Centre for Working Environment and Human Factors, Dortmund), for his very interesting feedbacks and for the exciting experiments we carried out together. I would also like to thank Prof. Angelo Cangelosi (Plymouth University, UK) and the rest of the committee for their interest in being part of the evaluation committee of my Ph.D. Thanks to Dr. Haris Dindo, Prof. Antonio Gentile, Prof. Marco La Cascia, Dr. Rosario Sorbello for their support during my B.Sc. and M.Sc. studies at the Computer Science Department of the University of Palermo (Italy).

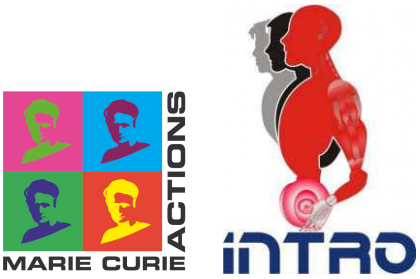
I would like to acknowledge all the participants of the experiments presented along this thesis. In particular, the participants of the experiment presented in Appendix B are Annika Dix, Lenka Dražanovà, Jovana Dačkovìc and Romy Frömer for their help with the statistical analysis and Levente Littvay for the latent growth curve analysis.

Thanks to all the people which made these three years in Berlin easier and funnier, in particular to Lenka, Fabio and Julienne, Giovanni and Eva, Julia, Matthias and Stefanie, Manuel, Saša, Raj, Francesca, Asif, Serena, Elvira, Barbara, to Alexandra Gründel and her German classes (and follow-ups), to Giacomo and our Swedish experience.

An enormous thank you to all my dear friends in Palermo. I counted down the days, before each time we met during the last three years. In particular, my dearest Totò and Francesca, Piero, Renzo, Sergio, Manlio, Luca, Girò, Lillo, Salvo and Adriana, Daniele,

Salvo Romano and the others. Please forgive me if I do not write all of your names down, but it would be a very long list.

Last but not least, I would like to express my gratitude to my wonderful parents Carlo and Aurelia, and to my siblings and their partners, Anna and Antonio, Giancarlo, Marina and Eliseo, Fabio and Julienne, for their unconditional support, confidence and help that I never lacked along my life. Undoubtedly, I could not have written this thesis without your presence. Thanks to my dear Lenka and to our daughter Alice, for our past and future great moments.



This work has been financed by the EU funded Initial Training Network (ITN) in the Marie-Curie People Programme (FP7): INTRO (INTeractive RObotics research network), grant agreement no.: 238486. http://introbotics.eu

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traditionally, robots played the role of powerful and fast machines in automated manufacturing processes. During the last decades, researchers spent big efforts on widening their range of application. The challenge has been to adopt artificial agents in different application domains than just factories: from service domains, such as surveillance, inspecting and renovating in hazardous environments, agriculture, firefighting, medicine, floor cleaning - to personal and social domains, such as entertainment, elderly and nursing care, autism therapy, rehabilitation (Dautenhahn, 2003).

The reasons why adopting robots in these contexts can be beneficial are manifold. For example, robot co-workers in hazardous scenarios could reduce the risk of human injuries. In the context of autism therapy, robots can play the role of social mediators facilitating social behaviour among children and among children and adults, or therapeutic playmates (Dautenhahn, 2003).

However, many factors are still slowing down the spread of robots outside of labs and factories. From a technical point of view, there are still several open challenges in the implementation of motor and cognitive skills in artificial agents. In fact, differently from biological systems, state-of-the-art robots are still not properly able to learn, adapt, react to unexpected circumstances, exhibit a proper level of intelligence and autonomously and safely operate in unconstrained and uncertain environments.

In dealing with these issues, roboticists broadened their research boundaries towards developmental psychology, cognitive science, neuroscience and even philosophy. Consequently, several new research fields emerged in robotics, such as *developmental robotics* that aims at building artificial intelligent systems inspired by theories on human development and at gaining new insights into the nature of intelligence (Lungarella et al., 2003). In fact, robots are increasingly used as research tools in comparative studies. On the one hand, roboticists can take inspiration from developmental psychology, cognitive sciences and neuroscience for the development of cognition in artificial systems. On the other hand, robots can be seen as test-beds for the theories in the mentioned fields (Dautenhahn, 2003).

In line with these trends, this thesis investigates mechanisms for the autonomous development of motor and cognitive capabilities in robots. In particular, it deals with some of the challenges related to the autonomous motor and mental development in artificial

systems. The *research questions* behind this work are:

1. What are the basic behavioural components an artificial agent should be provided with for being able to develop motor and cognitive capabilities?
2. How can an artificial agent represent and store the experience generated through such basic behaviours?
3. How can the acquired experience be reused and what computational processes are needed for generating basic cognitive skills out of it?

1.1 Approach

The approach adopted in this work can be summarised by the following three points.

1.1.1 Embodied Cognition

The assumption behind the studies presented in this work is that cognition is rooted in the bodily experience with the world. As it will be reviewed in Chapter 2, considering cognition as deeply intertwined with the embodiment of the individual is an account that dates back to the ancient Greek philosophy. However, whether cognition and body are separate substances or not has been a very debated topic over the following centuries. In its early stages, research in Artificial Intelligence was driven by a Dualist approach (mind and body as two separate entities), which resulted in producing artificial systems that under many aspects did not manage to deal with the challenges elicited in the previous section. Only few decades ago, the AI scientific community switched its attention back to an embodied view of cognition.

In robotics, many improvements have been introduced as a result of the application of the theories on embodied cognition. For example, (Pfeifer and Gómez, 2009) demonstrated that the morphology or the other physical characteristics of an embodied agent can take over some of the control processes in the context of locomotion, grasping or sensorimotor coordination.

Learning and interacting with the external world can be heavily influenced by the morphological properties of the agent. This thesis investigates sensorimotor interactions as a mean for the acquisition of experience in robots, where such interactions are resulting from the agents' bodily characteristics. The mechanisms presented in this work are not limited to a specific hardware platform, rather they are intended to be transferable to any artificial system equipped with sensory and motor modalities.

1.1.2 Simulation of experience as a computational process

Many of the robotic studies which applied theories of embodied cognition deal only with *online* aspects of cognition, that is with those processes that occur with the involvement of actual bodily experience, such as control of movements. In fact, cognition is not always an online phenomenon, since it can occur also in the absence of external stimuli.

Supported by evidence in the neurosciences and in behavioural science, theorists on grounded cognition suggested that mechanisms for mental simulation of sensorimotor experience could be a form of computation in the human brain that could account for some of those cognitive processes that occur *offline*, that is in absence of external stimuli (Barsalou, 2008). Mental simulations are intended as imaginary re-enactments of sensory and motor states that have been experienced in past interactions with the external world. Similar mechanisms are thought to be involved in mental processes such as mental imagery and in the functioning of the mirror neurons system (Rizzolatti and Craighero, 2004).

However, as it will be reviewed in the following chapters, very few robotic studies presented implementations of mental simulation processes. Undoubtedly, such mechanisms have still not been fully studied and exploited in the field of cognitive robotics. This thesis approaches the problem of implementing basic cognitive capabilities in robots by investigating the use of sensorimotor simulation processes.

1.1.3 Developmental learning

The aim of this work is not to advance the state-of-the-art in the implementation of robust and effective robot behaviours, rather to study how artificial embodied agents could acquire basic motor and cognitive capabilities by interacting with their environment as well as humans do during early developmental stages. This thesis adopts a developmental paradigm. Such an approach is not motivated by a mere interest in mimicking human development in artificial agents. Rather, studying human development could give insights in finding those basic behavioural components that may allow for the autonomous mental and motor development in artificial agents.

As argued before, embodiment is a crucial factor to take into account when implementing cognitive skills in artificial agents. However, defining models of robots' embodiment and their surrounding world *a priori* should be avoided. The risk is to stumble across problems such as robot behaviours lacking of adaptability and of capability to react to unexpected circumstances. A branch of the robotics community known as *developmental robotics* investigates techniques for motor and cognitive development in artificial systems. The aim is to provide artificial agents with mechanisms based on long-term interactions with the physical and social environment, through which they can develop increasingly more complex motor and cognitive capabilities and become more autonomous, adaptable and social (Lungarella et al., 2003).

This thesis investigates some of the basic mechanisms that can provide an artificial agent with a mean for the acquisition of experience through the interaction with its surrounding world. Such mechanisms, such as exploration behaviours, mapping of multimodal information, attentive behaviours, are inspired on human developmental studies. Moreover, they are thought to be prerequisites for sensorimotor and social development.

1.2 Contributions

The contributions of this thesis are manifold. Firstly, it deepens the study on multi-modal frameworks for representing sensorimotor behaviours in artificial agents. In particular, it

adopts the established internal models framework (Wolpert and Kawato, 1998), namely inverse and forward models, for coding sensorimotor experience acquired through self-exploration in a humanoid robot. I believe that such a framework has not been fully studied and exploited in robotics. This thesis contributes with a deeper investigation of internal models by making the least assumptions possible in the construction of them. In particular, inverse and forward models have been trained with low-level sensory and motor data generated by the robot through exploration behaviours, learning by demonstration and kinesthetic teaching. In addition, the adoption of the internal models framework is motivated by its capability to generate simulations of sensorimotor loops. As it will be shown in the following chapters, internal simulations have been used as computational processes behind the implementation of basic cognitive capabilities in a humanoid robot.

Besides the adoption of internal models and the implementation of internal simulation mechanisms, this thesis presents implementations of basic cognitive capabilities in robots following a developmental paradigm. Figure 1.1 illustrates the skills that have been implemented and that will be presented along this thesis. The skills are elicited along two developmental timelines (*sensorimotor* and *social*) and ordered by complexity.

The sensorimotor developmental line takes inspiration from the earliest sensorimotor stages depicted in Piaget's theory of cognitive development (Piaget, 1983). Piaget elicited a list of stages a child goes through during the first months of her/his life. The common ground of those stages is the exploration of motor behaviours with the so-called circular reactions, that is, processes of repetition of movements that the child finds pleasurable. Through such rehearsals, infants acquire governance and coordination of their motor capabilities. Around the age of 8-12 months, children begin to understand the objects they are surrounded by, to recognise their properties and the sensorimotor contingencies they determine, and to play with them.

In this work, a basic behavioural component has been identified, namely body babbling (in Figure 1.1, labelled as the zero-point (**ZP**) of the developmental timeline), which allows an artificial agent to autonomously acquire sensorimotor experience by self-exploration. The sensorimotor developmental line of Figure 1.1 contains incrementally more complex motor behaviours that have been acquired through self-exploration mechanisms. Mental simulation mechanisms have been used in implementing basic cognitive skills out of the acquired sensorimotor experience.

In particular, the experiments related to the sensorimotor (SM) developmental line that will be presented along this thesis are:

SM1. Generation of body maps through motor babbling. Body maps model the relationships between muscles, or effectors, activations and sensory perceptions. As well as infants, an artificial agent can learn them over time. Through exploration behaviours, a robot can generate sensorimotor data and model it as multi-modal maps.

SM2. Characterization of self-produced movements through the use of the internal models framework. This experiment is used as a basis for the study on self-other distinction (**SO4**).

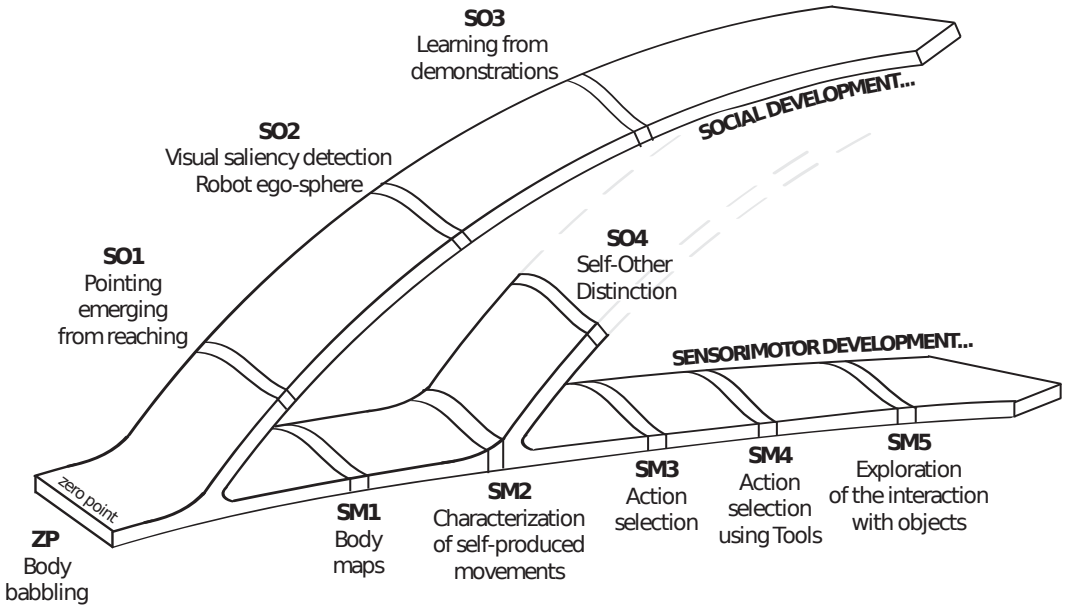


Figure 1.1: An illustration of the robot capabilities implemented in this thesis. The bottom-left corner represents the origin of the two developmental timelines (*sensorimotor* and *social*). The origin of the two lines indicates the most basic behaviour, namely body babbling (or the *zero point*, **ZP**), that has been implemented in the robot as a mean for the acquisition of most of the other more complex motor skills. As it will be discussed in Section 3.1, body babbling is a self-exploration behaviour exhibited by infants and that is thought to be the zero-point in the development of motor capabilities. Section 3.2 illustrates how a humanoid robot equipped with self-exploration behaviours can learn its body maps (**SM1**). Section 4.3.3 studies how self-produced movements can be characterised through the use of internal models (**SM2**). As a result of this study, an account on self-other distinction based on mental simulations of hand trajectories is given (**SO4**). Internal models and internal simulations are also used for representing arm movements and for implementing action selection capabilities (**SM3**), as described in Section 4.3.1. In addition, Section 4.3.2 illustrates how tool-use can affect the body schema and how self-exploration behaviours allot the robot to deal with the new body configuration (**SM4**). Thus, experiments on the recognition of motor actions involving objects will be presented, where the robot acquired such motor capabilities by self-exploration (**SM5**, Section 4.3.5) or by observing a human demonstrator (**SO3** on the social developmental line, Section 4.3.5). The first two behaviours on the social developmental line have been implemented as prerequisites of joint attention, the capability that humans have to share the attention with other individuals. In particular, Section 3.3 shows how the humanoid robot Nao can exhibit proto-imperative pointing gestures as a result of failed grasping actions (**SO1**). In addition, an implementation of visual saliency detection skills and short-memory system based on a robot ego-sphere (**SO2**) will be presented in Section 3.4. As it will be discussed in Appendix B, the adoption of attention mechanisms, such as visual saliency detection and attention manipulation, can improve human-robot interaction.

SM3. Acquisition of sensorimotor schemes such as arm movements through motor babbling. The internal models framework is adopted for coding the acquired sensorimotor schemes. Internal simulations mechanisms are used for implementing action selection capabilities.

SM4. Extension of SM3 with the use of tools which modify the morphology of the robot. Self-exploration behaviours allot the robot to deal with the new body configuration.

SM5. Acquisition of sensorimotor behaviours involving interactions with objects through motor babbling. The internal models framework is adopted for coding the acquired sensorimotor behaviours. Internal simulation mechanisms are used for implementing motor behaviours recognition capabilities.

The second line depicted in Figure 1.1 is concerned with the development of social (SO) skills in artificial agents. This thesis presents an implementation of some of the behavioural prerequisites of joint attention capabilities and of imitation learning, in particular:

SO1. Imperative pointing gestures as a result of failed reaching/grasping behaviours, which have been acquired through motor babbling (**ZP**). Imperative pointing is one of the first step in the development of attention manipulation capabilities in robots.

SO2. Visual saliency detection and short-memory system based on an ego-sphere. A side study presented in Appendix B will show how behaviours based on attention mechanisms, such as pointing gestures or exploration based on visual saliency detection, affects human-robot interaction.

SO3. Acquisition of sensorimotor behaviours involving interactions with objects through observing a human demonstrator and implementation of motor behaviours recognition capabilities.

SO4. Self-other distinction capabilities based on classifications of the velocity profiles of observed hand trajectories. The capability to distinguish between the Self and the Other is a fundamental prerequisite for social interaction and social development. In line with studies on the visual perception of biological motion and self-other distinction in humans, this thesis explores the possibility that also robotic motion can have specific properties that makes it different from other types of motion.

The separation between sensorimotor and social development as depicted in Figure 1.1 is made for the purpose of clarification in eliciting the motor and cognitive capabilities presented along this thesis. The aim here is not to take a position in the debate of the Piagetian/Vygotskyan accounts on cognitive development (details in Chapter 3, nor to say that children follow two separate developmental paths at the same time. Rather, the aim is to identify those basic skills that could allow for further development of motor and cognitive skills in robots through the sensorimotor and social contexts.

It should be pointed out that the experiments have not been carried out in a one-shot long term learning fashion, with each skill built on top of each other and resulting from the

previous ones. Rather, most of the experiments have been carried out separately, although they share common procedures: gathering sensorimotor experience through exploration behaviours and reusing such experience in mental simulations processes for implementing cognitive capabilities. As it will be demonstrated in the following chapters, these processes can be used as common building blocks for the development of cognition in robots.

1.3 Structure of the Thesis

The rest of this thesis is structured as follows.

Chapter 2 introduces theories of embodied cognition and studies on processes of mental simulations. In particular, Section 2.1 reviews the history of the study of human cognition within artificial intelligence, from Cartesian Dualism to embodied cognition theories. Thus, Section 2.2 reviews philosophical and scientific studies on mental imagery and simulation of experience together with a discussion on theories of grounded cognition.

Chapter 3 focuses on human development and on developmental robotics. The aim is to identify the basic behavioural components that a robot should be equipped with for the autonomous development of cognition. Thus, relevant studies in the field of human motor and cognitive development and of developmental robotics are reviewed. Section 3.1 deepens the study on motor development by presenting an implementation of self-exploration behaviour on the humanoid robot Aldebaran Nao. In Section 3.2, a mechanism based on sensorimotor coordination will be presented for generating multimodal information and for building body maps. Appendix A presents a side study on the implementation of random movement strategies for self-exploration in robots. The rest of Chapter 3 is concerned with the implementation of some of the prerequisites for joint attention. In particular, Section 3.3 investigates how proto-imperative pointing gestures can emerge from failed reaching/grasping behaviours. Section 3.4 presents an implementation of visual saliency detection mechanisms and of a short-memory system based on a robot ego-sphere. Appendix B presents a side study in which human participants evaluated the effect of saliency detection and attention manipulation mechanisms in human-robot interaction.

Chapter 4 begins with a review of the studies supporting the existence of simulation mechanisms in human cognitive processes. Thus, it describes the framework of the internal models, namely inverse and forward models, which is adopted for coding sensorimotor skills acquired through exploration behaviours in robots. This framework can provide a robot with multi-modal representations of motor behaviours and it can be adopted as a tool for simulation of experience. Chapter 4 also presents a review of some of the most well-known architectures and implementations in the area in the search to identify and try to fill some of the gaps in their study. Then, Chapter 4 presents the experiments listed in Figure 1.1 that share the following procedures: sensorimotor learning; internal models for coding sensorimotor experience; simulation processes for implementing basic cognitive capabilities. In particular, Section 4.3.1 shows an implementation of action selection capabilities, in which the robot is capable of selecting one of the two arms to use for reaching a desired target point, based on its past sensorimotor experience and on simulating the outcome of its motor actions. Section 4.3.2 present an extension of the

previous experiment, in which the robot's morphology has been extended with the use of a tool. Section 4.3.4 presents an account on self-other distinction capabilities based on mental simulations of hand trajectories, which are characterised in Section 4.3.3. Section 4.3.5 introduces experiments where the robot learns basic interactions with objects and stores the gathered sensorimotor experience into an internal model framework. Two learning paradigms have been tested, as described in Section 4.3.5: self-exploration or learning from demonstrations. Section 4.3.5 shows an experiment on the recognition of motor behaviours involving interaction with objects, where such behaviours have been learned through self-exploration and through observing human demonstrations. Section 4.3.5 also shows how internal simulations can be used in detecting the target object of a motor action.

Finally, Chapter 5 summarises the contributions of the thesis and concludes with an outlook of possible applications and extensions.

1.4 Publications

In accordance with the doctorate regulations (Promotionsordnung) of the Mathematisch-Naturwissenschaftliche Fakultät II at the Humboldt-Universität zu Berlin, Section 6, Article 2.b, this thesis is based on the works presented in the following articles:[1]

Journal papers

- **Schillaci, G.**, Bodiroza, S., Hafner, V.V., Evaluating the Effect of Saliency Detection and Attention Manipulation in Human-Robot Interaction, International Journal of Social Robotics, DOI: 10.1007/s12369-012-0174-7. November, 2012.

- Hafner, V.V., **Schillaci, G.**, From Field of View to Field of Reach Could Pointing Emerge from the Development of Grasping?, IEEE Conference on Development and Learning and Epigenetic Robotics (IEEE ICDL-EPIROB 2011). Conference abstract published in Frontiers in Computational Neuroscience. August, 2011.

Book Chapters

- **Schillaci, G.**, Lara, B., Hafner, V. V., Internal Simulations for Behaviour Selection and Recognition, in Lecture Notes in Computer Science, pp. 148-160, Springer 2012, ISBN 978-3-642-34013-0, Albert Ali Salah, Javier Ruiz-del-Solar, Çetin Meriçli, Pierre-Yves Oudeyer (Eds.): Proceedings of Human Behavior Understanding - Third International Workshop, IROS-HBU 2012, Vilamoura, Portugal. October 7, 2012.

[1] My contribution in the design and implementation of the experiment is the major part where mentioned as first author. In the integration work presented in (Kozlov et al., 2013), I provided only the source code for executing motor babbling behaviours, for training inverse models (as neural networks) and for executing reaching actions, which have been integrated in the mobile robot platform and tested by the other authors.

In (Bodiroža et al., 2011), all the authors actively participated in the design and implementation of the experiments. The tessellated ego-sphere and the related habituation, inhibition and forgetting mechanisms have been implemented by Saša Bodiroža.

Conference Proceedings

- Kozlov, A., Gancet, J., Letier, P., **Schillaci, G.**, Hafner, V.V., Fonooni, B., Nevatia, Y., Hellstrom, T., Development of a search and rescue field robotic assistant. Proceedings of the 11th IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR2013), pp.1–5, Linkoping, Sweden. October 2013.
- **Schillaci, G.**, Hafner, V. V., Lara, B., Grosjean, M., Is That Me? Sensorimotor Learning and Self-Other Distinction in Robotics, Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013), Tokyo, Japan. March, 2013.
- **Schillaci, G.**, Hafner, V. V., Lara, B., Coupled Inverse-Forward Models for Action Execution Leading to Tool-Use in a Humanoid Robot, Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012), Boston, USA. March, 2012.
- **Schillaci, G.**, Lara, B., Hafner, V. V., Internal Simulation of the Sensorimotor Loop in Action Execution and Recognition, In Proceedings of the 5th International Conference on Cognitive Systems (CogSys2012), p. 105, TU Vienna, Austria. February, 2012.
- Bodiroža, S., **Schillaci, G.** and Hafner, V.V., Robot Ego-sphere: An Approach for Saliency Detection and Attention Manipulation in Humanoid Robots for Intuitive Interaction, in Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots, pp. 689-694, Bled, Slovenia, ISBN: 978-1-61284-867-9. October, 2011.
- **Schillaci, G.** and Hafner, V.V., Random Movement Strategies in Self-Exploration for a Humanoid Robot, in Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2011), pp. 245-246. March, 2011.
- **Schillaci, G.** and Hafner, V.V., Prerequisites for intuitive interaction - on the example of humanoid motor babbling, in Proceedings of the Workshop on The role of expectations in intuitive human-robot interaction (at HRI 2011), pp. 23-27. March, 2011.

Journal paper in preparation

- **Schillaci, G.**, Rodriguez, D., Hafner, V.V., Lara, B., Internal Models Encoding Low Level Sensorimotor Representations in Humanoid Robots.

Posters not in proceedings

- **Schillaci, G.**, Hafner, V. V., Lara, B., I Know How I Would Do It. Internal Simulations of Sensorimotor Experience. Robotics Science and Systems Conference, Workshop in Active learning in Robotics: Exploration, Curiosity, and Interaction. Berlin, Germany. June, 2013.

- **Schillaci, G.**, Hafner, V. V., Lara, B., Sensorimotor Loop Simulations as a Prerequisite for Imitation. Experiments with a Humanoid Robot, International Conference on Knowledge through Interaction: How children learn about self, others and objects, Heidelberg, Germany. March, 2012.

Publications not directly relevant to this thesis

- Dindo, H., **Schillaci, G.**. An Adaptive Probabilistic Approach to Goal-Level Imitation Learning, in 2010 IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS 2010), Taipei, Taiwan. October, 2010.

- Dindo, H., **Schillaci, G.**. An Adaptive Probabilistic Graphical Model for Representing Skills in PbD Settings, in 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), Osaka, Japan. March, 2010

Chapter 2

Embodied cognition and Mental Imagery

As introduced in the previous chapter, this thesis is based on the assumption that cognition is rooted in the bodily experience with the world. As a result, cognitive processes must be implemented in artificial agents as strongly intertwined with their bodily characteristics. This account is recent in the field of artificial intelligence. Section 2.1 reviews the history of the study of cognition with artificial intelligence, pointing out theories of embodied cognition and the relevant studies in behavioural sciences.

The second assumption behind this thesis is that mental simulations of experience could be behind the functioning of basic cognitive processes in humans. Such processes can serve as computational mechanisms behind the development of cognition in robots. Section 2.2 reviews the history of the study on mental imagery and on mechanisms of internal simulations, where mental simulations are intended as sensory or motor re-enactments or quasi-perceptual experiences that occur in the absence of external stimuli.

2.1 Body and Mind

Until a few decades ago, the mainstream of the research in Artificial Intelligence was logic and problem solving. Researchers focused their efforts in trying to reproduce human intelligence in artificial systems, based on the assumption that human cognition was a complex manipulation process of mental representations of the external world. Several achievements have been attained following this paradigm, such as programming computer chess players capable of winning matches against human world champions. However, such an approach based on abstract representations of the external world did not manage to produce artificial systems capable to react and adapt to unexpected environmental circumstances.

Since its early ages, research on artificial intelligence has been influenced to a certain extent by the contemporary philosophical positions regarding the body and mind problem: are material and mental belonging to the same realm or to two different ones? Such a fundamental question has been one of the most debated philosophical topics. John Mingers

distinguishes three periods that differ in terms of the assumptions about the body and mind problem within artificial intelligence (Mingers, 2001), which consequently influenced the way artificial systems were developed: the first is based on the disembodied Cartesian Dualism; the second is inspired by the work of Heidegger on practical activity in the world; the third is based on the view of cognition as inherently embodied, which can be in part related to the phenomenology[1] of Merleau-Ponty.

The first period is based on the assumption that body and mind are belonging to two different substances. Often referred in the field of Philosophy of Mind as *Dualism*, the roots of this assumption can be found already in Greek philosophy (see (Robinson, 2012) for a review on the history of Dualism). For instance, Plato believed that the ephemeral physical bodies are only imperfect copies of the eternal Forms, the true substances. Similarly, Aristotle argued in *De Anima* that the intellect is immaterial and it differs from other faculties in not having a bodily organ. Even most theologies consider immortal souls as existing in a realm that is independent from the material world.

However, it was only in the 17th century that the first systematic account of Dualism was given. Descartes proposed what is nowadays known as Cartesian Dualism, according to which mind and brain belong to two different kinds of substances: mental and material. In the following centuries, many philosophers supported Dualism, including Edmund Husserl, who reinforced Descartes' ideas with the formulation of the *trascendental ego* and the *phenomenological epoché*(Mingers, 2001).

In the cognitive science, Cartesian Dualism formed the basis of the Cognitivist hypothesis. Cognitivism was a movement born in response to Behaviourism in the 1950s. Influenced by the works of Newell, Simon, Chomsky and Fodor, Cognitivism viewed cognition as manipulations of symbolic representations of the world. According to this view, cognition occurs by taking in information provided by the environment, forming this into representations and then processing them to provide logical responses, as well as an information-processing machine would do (Mingers, 2001).

In 1956, McCarthy, Minsky, Shannon and Nat Rochester held a conference in Dartmouth which is considered to have given birth to artificial intelligence (McCarthy et al., 2006). Cartesian representationalism was the main paradigm in their discussions on human cognition and its reproduction into artificial machines (Miller, 2003). For some decades ahead, the cognitivist approach dominated the research in artificial intelligence and influenced the way computational models of human intelligence were designed. However, adopting the paradigm of the artificial brain that processes symbols which are related together to form representations of the world outside was not successful. Evidences can be found in the field of robotics, where this approach hardly can be adopted for mimicking even the basic human abilities, such as perception, physical manipulation and speech (Mingers, 2001).

According to Mingers, the second period in the history of the study of the nature of human cognition within artificial intelligence started around the 1970s, when Hubert Dreyfus strongly criticised what John Haugeland named *Good Old-Fashioned Artificial Intelligence*

[1]Phenomenology is the study of structures of consciousness as experienced from the first-person point of view (Smith, 2011).

(GOFAI), for implicating the failure and the stagnation of artificial intelligence (Dreyfus, 1972). Dreyfus blamed GOFAI researchers in their attempt to model intelligence as manipulations of symbolic representations of the world. In arguing that mimicking human intelligence requires bodily agents *in the world*, Dreyfus was one of the first researchers to introduce the ideas of the German philosopher Martin Heidegger into the AI community.

As discussed in (Megill, 2003), Heidegger posed a strong critique against the Dualist account of human cognition with the notion of *Being-in-the-World*. According to Heidegger's Being-in-the-World notion, subject and object (mental and material) are connected in a way that it is not possible to differentiate the two: "Self and world belong together in a single entity: Dasein". This entity is immersed in the world and it is most of the time engaged in tasks, or *practical activity*. During this engagement in the task, *Dasein* makes use of objects (or *equipments*) but only when they can have an unreflective relevance for the accomplishment of the task (Megill, 2003). In demonstrating this, Heidegger made the example of a carpenter hammering a nail. During this practical task, *Dasein* is non-consciously directed at the task and at steps to perform for accomplishing it; instead, it uses the hammer only as a mere equipment, without having to think at the whole concept of the object *hammer*. Only in case of a malfunctioning, *Dasein* interrupts the unreflective immersion in the task. This would result in the emergence of a conscious subject which reflects upon the external world in a theoretical manner (Megill, 2003). However, practical activity can exist without theoretical cognition, but not vice versa. Nonetheless, the notion of *Dasein* states a departure from the Dualist view of conscious subject separated from the external world.

Heiddeger's ideas, through Dreyfus' critiques, started the breaking down of the foundations of traditional artificial intelligence. It was the time of Maturana and Varela's proposal of *structural coupling* between organism and environment (Maturana and Varela, 1987) or the language/action approach of Winograd and Flores (Winograd and Flores, 1987), according to which cognition is not seen any more as an isolated mental function, but a normal everyday activity, and knowledge does not consist of representations, rather, we structure and restructure the world as we co-ordinate our purposeful activities (Mingers, 2001).

In the late 1980s - early 1990s, motivated by a new philosophical interest on embodiment (such as on the work of Heidegger and Merleau-Ponty[2]), a new era known as post-cognitivism started to flourish in the cognitive sciences. The post-cognitivist ideas rapidly spread into the artificial intelligence community bringing the message of the importance of the body in cognition.

In (Wilson and Foglia, 2011), the authors cite three landmark publications representative of the earliest post-cognitivist ideas: *Metaphors We Live By* (1980) by George Lakoff and Mark Johnson, *The Embodied Mind* (1991) by Francisco Varela, Evan Thompson, and Eleanor Rosch, and *Being There: Putting Mind, World, and Body Back Together* (1997)

[2]In *The Phenomenology of Perception*, Merleau-Ponty viewed human subjectivity, the *incarnate subject*, as an embodied phenomenon. The brain would be part of a larger system, the nervous system and the entire body, and it is to this larger system that we must turn if we are to understand intelligent behavior (Loren and Dietrich, 1997). Merleau-Ponty's work was one of the most influential during the early stages of post-cognitivism.

by Andy Clark.

In *Metaphors We Live By*, Lakoff and Johnson argued that cognitive processes, such as those concerning space and time, are both expressed and influenced by metaphors (for example, *mind is a computer*, *mind is a container*, *ideas are objects*, etc.). Metaphors are culturally defined and based on personal experiences and they shape our perceptions and actions. Cognition is embodied because our experiences and metaphors are shaped by our bodies which mediate between us and the external world.

In (Varela et al., 1992), Varela, Thompson and Rosch introduced the concept of *enaction*: the interactions between the body, its sensorimotor circuit and the environment determine the way the world is experienced. Cognitive agents are *living bodies* situated in the environment and knowledge emerges through the bodily engagement with it.

In *Being There* (1997), Andy Clark provided an integrative framework for the emerging works of the late 1980s and the early 1990s on embodiment in the cognitive sciences and in robotics (Wilson and Foglia, 2011). According to Clark, biological brains are control systems for biological bodies and cognition is a situated activity that takes place in the context of task-relevant inputs and outputs.

In the last two decades, many studies in behavioural sciences, social psychology on attitudes, emotion and social perception supported the idea that body is closely tied with cognition. For instance, the famous experiment of Strack and colleagues demonstrated that people's facial activity influences their affective responses (Strack et al., 1988). Subjects were holding a pen in their mouth in ways that either inhibited or facilitated the muscles typically associated with smiling without requiring subjects to pose in a smiling face. The authors found that subjects reported more intense humour responses when cartoons were presented under facilitating conditions than under inhibiting conditions.

Facial expressions have been found to influence affective self-reports. In (Laird, 1974), participants were required to smile and frown without awareness of the nature of their expressions. Subjects reported feeling more angry when frowning and more happy when smiling.

(Wells and Petty, 1980) demonstrated that overt movement can influence cognitive activities. Subjects who believed that they were testing headphone sets engaged in either vertical, horizontal, or non-instructed head movements while listening to a simulated radio broadcast. Subjects in the vertical head movement conditions agreed with the editorial content of the radio broadcast more than did those in the horizontal head-movement conditions (Wells and Petty, 1980). One of the explanations is that nodding movements have been positively associated with cognitive activity in the past, whether horizontal head movement has been associated with negative ones.

In the field of psychology of perception, already in the 1950s, J.J. Gibson argued that perceptual learning should not be reduced to stimulus-response theory. Rather, perception should be viewed as an aspect of the individual's interaction with the environment. As he claimed in his theory of affordances, the perceptual and motor systems are interdependent and the way how we perceive the world is shaped by object possibilities for action (Gibson, 1977).

In the artificial intelligence community, the embodied cognition proposal gained such a broad support that researchers started referring to *embodied artificial intelligence*. How-

ever, as noted by Margaret Wilson in (Wilson, 2002), there is a diversity in the claims, for which it is important to be cautious to scale up such principles to explain the whole human cognition. In pointing out six viewpoints (*cognition is situated*, *cognition is time pressured*, *cognitive work is off-load onto the environment*, *environment is part of the cognitive system*, *cognition is for action*, *off-line cognition is body based*), she reaches the conclusion that it is beneficial distinguishing between on-line aspects of embodied cognition and off-line aspects. For example, in the case of the claim *cognition is situated*, one could note that portions of human cognitive processes are excluded. By definition, situated cognition involves interaction with the things that the cognitive activity is about. Thus, she concludes, any cognitive activity that take place *off-line* (such as planning, remembering, day-dreaming) is not situated (Wilson, 2002).

According to Wilson, the claim *off-line cognition is body based* received the least attention in the literature on embodied cognition. Off-line aspects of embodied cognition refer to those cognitive activities in which imaginary (or distant in time and space) sensory and motor information are used in mental tasks. As discussed in the next section, evidences suggest that processes such as sensorimotor simulations of external situations may be at the basis of several phenomena in human cognition.

2.2 Mental Imagery

Imagining a picture or the smell of something is a familiar mental phenomenon that most of us experience almost daily. Known in the philosophical and scientific literature as *mental imagery*, this phenomenon has been defined as a quasi-perceptual experience which resembles perceptual experience but occurs in absence of external stimuli (Thomas, 2013). As recognised by contemporary cognitive scientists, mental imagery does not refer only to visual imagination but to *imagining* in any sensory modalities. In fact, several studies can be found in the literature on auditory imagery, olfactory imagery, kinaesthetic imagery, and so on.

What the nature is of this mental phenomenon has always been a very debated topic ((Thomas, 2013) provides a more comprehensive review of the literature on mental imagery). Not surprisingly, studies on mental imagery can be found already in the Greek philosophy. In *De Anima*, Aristotle saw mental images, residues of actual impressions or *phantasmata* as playing a central role in human cognition, for example in memory.

At the beginning of the 20th century, behaviourists such as J. B. Watson were skeptic about the psychological importance of imagery. Behaviourists believed that psychology must have dealt only with observable behaviours of people and animals, not with unobservable introspective events. Therefore, mental imagery was reputed as not being sufficiently scientific (Watson, 1913), since no rigorous experimental method was proposed to demonstrate it.

Only after the 1960s, perhaps motivated by studies on hallucinogenic drugs, perceptual problems[3] (Holt, 1964) and by the development of electroencephalography, mental imagery

[3]Studies on people whose work required them to remain perceptually alert (such as radar operators and jet pilots) while observing barely changing visual stimuli over long periods of time, reports subjects

gained new attention (Thomas, 2013). It was the time of the famous works by Shepard and Metzler on mental rotations (Shepard and Metzler, 1971) and by Kosslyn on mental scanning of visual images (Kosslyn et al., 1978).

Cooper and Podgorny provided a demonstration of the analog nature of mental rotation tasks (Cooper and Podgorny, 1975). In particular, they instructed participants to imagine a mental rotation of a two-dimensional shape. At some point during the mental rotation, a test shape was presented at a certain orientation, and the subject was required to determine as rapidly as possible whether the test shape was the same as the originally designated shape or was its mirror image. When the test shape was presented in the expected orientation, the reaction time was short and constant, regardless of the angular departure of that orientation from a previously trained position (Cooper and Podgorny, 1975). This finding suggested that during a mental rotation the internal process passes through a trajectory of intermediate states which have a one-to-one correspondence to the intermediate stages in an external rotation.

Around the 1980s, Kosslyn proposed what is nowadays known as the *analog* or *quasi-pictorial* account of the nature of mental imagery (Kosslyn, 1980), opposed to the *propositional* one supported by (Pylyshyn, 1977). The analog/propositional debate is concerned with the dispute in cognitive science about the nature of the representational format of visual mental imagery (Thomas, 2013). Scientists supporting the analog side considered mental imagery experiences as pictures with the same spatial properties that real pictures have. In a sense, according to the analog view, representations have the same structure as the thing represented. On the other side, Pylyshyn and the *new cognitivists* (as defined by (Barsalou, 2008)) defended the propositional nature of visual mental images, assuming that mental representations are more like linguistics descriptions of visual scenes (Thomas, 2013).

A third account on the nature of mental imagery was based on the *enactive* theory of perception proposed by J.J. Gibson (Gibson, 1966). The *enactive* theory (or perceptual activity theory) of imagery was based on the idea that perception is not passive, but is a form of action (Thomas, 1999). Organisms actively explore the environment, seeking out what they search based on the sensory stimuli they perceive. Imagery is thus experienced when one persists in acting out the seeking of some particular information, even if the information is not expected to be there (Thomas, 2013). In other words, mental imagery is the enactment of the visual perception of what is imagined, that is the execution of those subconscious processes that guide the recognition of an object during visual perception (e.g. which saccades are made to recognise a given object as that object) in absence of the object itself (Sima and Freksa, 2012).

However, it was only with the rise of grounded cognition theories (for a review, see (Barsalou, 2008) that the enactive view of mental imagery gained attention in the artificial intelligence community. Theories of grounded cognition proposed that knowledge is represented by modal representations and imagery. Researchers in grounded cognition criticised the traditional cognitivist approach for failing to explain how cognition interfaces

experiencing vivid and intrusive mental imagery when they were deprived of such visual perception in the laboratory.

with perception and action, for adopting formalisms based on amodal symbols, although little empirical evidence supported their presence in cognition, and for any lacking in explaining where the brain stores amodal symbols and how amodal symbols could be consistent with neural principles of computation (Barsalou, 2008).

As suggested by grounded cognition theories, modal simulations, such as recreations of perceptual, motor and introspective states, could account for the off-line characteristics of cognition, in which internal simulations of sensorimotor cycles are executed. Sensorimotor simulation processes have been found to be capable of modelling several behaviours and characteristics of the brain. Chapter 4 of this thesis reviews related studies.

Simulation processes have not been fully studied and exploited in robotics. As it will be presented in the following chapters, this thesis adopts the internal simulation paradigm for equipping robots with basic cognitive skills.

Chapter 3

Development of cognition

This chapter deals with the first research question that has been addressed at the beginning of this thesis: "What are the basic behavioural components an artificial agent should be provided with for being able to develop motor and cognitive capabilities?".

As argued in the previous chapter, embodiment is a crucial factor to take into account when implementing cognitive skills in artificial agents. However, defining models of robots' embodiment and their surrounding world *a priori* should be avoided. The risk is to stumble across problems such as robot behaviours lacking of adaptability and of capability to react to unexpected circumstances.

In fact, biological systems are not innately skilful in governing their body. It is reasonable to think that they do not possess innate models of their corporeality. Although instincts and primitive reflexes exist as in human and animal behaviours, complex motor and cognitive capabilities are developed over time. For instance, newborns exhibit primitives reflexes in response to certain stimuli. Examples are the rooting reflex that assists in the act of breastfeeding, the parachute reflex that protects from falls and the grasp reflex[1]. However, such mechanisms disappear along the developmental process, when they are substituted by more complex behaviours.

Since prenatal development, foetuses practice control of movements, a capability that will become the basis for the interaction with the outside world. Zoia et al. report a kinematic analysis to understand the movement dynamics of foetuses (Zoia et al., 2007). Analysing hand to mouth and hand to eye movements they observed that up to the gestational age of 18 weeks there was no evidence of coordinated kinematic patterns. Movements such as reaching were inaccurate and showed poor control of the hand trajectory with characteristics jerky and zigzag movements (Zoia et al., 2007). However, around the 22nd week of gestation, foetuses perform movements that show kinematic patterns with

[1]The rooting reflex can be stimulated by stroking the cheek of the infant, resulting in the baby turning its head towards the stimulus and making sucking movements with the mouth. The parachute reflex produces an extension of the arms of the baby and it occurs when rotating the body of the infant quickly from an upright position to a facing-down one, as if simulating a fall. The grasp reflex emerges already around the 11th week of gestation. The infant closes the hand around what touches its palm. Touching an infant's palm and trying to remove the finger causes the grip to tighten. Source: Infant reflexes, A.D.A.M. multimedia Encyclopedia. PennState Hershey, Milton S. Hershey Medical Center. http://pennstatehershey.adam.com

acceleration and deceleration phases apparently planned according to the size and to the delicacy of the target (facial parts, such as mouth or eyes). This finding supports the idea that control of movement is a capability that is acquired and refined over time through exploration behaviours, already during prenatal stages.

In developmental psychology, Jean Piaget suggested that the capability to coordinate sensorimotor interactions with the world is acquired during the first 2 years of infancy, a period that he named *sensorimotor stage*. As argued in (Piaget, 1983), infants would follow a sequence of developmental stages, starting from the sensorimotor one up to more complex social stages. In the sensorimotor stage, development would proceed along substages, whose common characteristic is the exploration of motor behaviours. Defined by Piaget as *circular reactions*, such behaviours consist of repetitions of movements that the child finds pleasurable. Through such rehearsals, infants acquire governance and coordination of those motor capabilities (such as reaching an object) that will enable them, subsequently, to explore the interactions with objects and with people (Sheldon, 2012).

However, Piaget has been criticised for viewing children as *solitary thinkers*, that is, for giving a central role only to the child's own activities in the building up of cognitive capabilities in early developmental stages (Dautenhahn and Billard, 1999). In fact, according to his view, the social context was only assisting the early stages of cognitive development. In particular, Piaget sustained that sensorimotor and intellectual development precedes social learning.

Whether intellectual development precedes social learning, or if they occur at the same time, has been a very debated topic in developmental psychology. Similarly to Piaget, the Russian researcher Lev Vygotsky also proposed that children develop in stages, but he emphasised the role of the social and cultural context in the cognitive development. Although having an active role in the learning process, children would be *scaffolded*, or supported, by their caregivers. In this sense, toddlers are not just solitary thinkers, rather they learn in a social context where caregivers scaffold their interaction with the world through the help of language. In particular, Vygotsky defined the Zone of Proximal Development where learners would complete tasks with the guidance of an expert. The Zone of Proximal Development lays between what the learner can do without any help and what the learner cannot do without guidance, during the developmental process.

Thus, according to Vygotsky, higher mental functions would be grounded on social interactions. Social interaction would shape and constrain biological factors, such as those that constitute embodiment, during a long ontogenetical period, after which advanced cognitive abilities would emerge (Lindblom and Ziemke, 2002).

Although this topic is still under debate in the developmental psychology community, it is clear that both social and non-social interaction skills are essential means for the acquisition of knowledge and for the development of cognition in artificial agents.

A branch of the robotics research known as *developmental robotics* investigates techniques for motor and cognitive development in artificial systems. The aim is to provide artificial agents with mechanisms based on long-term interactions with the physical and social environment, through which they can develop increasingly more complex motor and cognitive capabilities and become more autonomous, adaptable and social (Lungarella

et al., 2003). In addition, as pointed out in (Berthouze and Metta, 2005), studying cognition from a developmental point of view and applying theories into robots could provide deeper insights into the adult manifestation of cognitive skills.

(Lungarella et al., 2003) reviews studies on developmental robotics by collecting them into different areas of interest, including development of individual sensorimotor control, non-social interaction and socially-oriented interaction.

(Metta, 2000) investigated sensorimotor development as a model of learning and adaptation from a neuroscience and robotics perspective. He demonstrated how a twelve degrees of freedom humanoid robot (Babybot) acquires orienting and reaching behaviors following a developmental paradigm. Inspired by biological development of visuo-motor coordination, he implemented an adaptive control system for the robot that follows developmental stages, starting from a "plant" mostly driven by reflexes, and steering through phases where the cortex begins to influence sub-cortical structures (Metta, 2000). "At birth", the system is able to move the eyes only. Control, at that stage, is a mixture of random and goal-directed movements, such as reflexive behavior simulating basic muscular synergies and spinal reflexes. Thus, the development proceeds with the acquisition of closed loop gains, reflex-like modules controlling the arm sub-system, acquisition of eye-head coordination and of head-arm coordination map.

(Asada et al., 2009) raised a criticism against previous developmental robotics approaches related to the fact that they often explicitly implemented control structures derived from the designer's understanding of the robot's physics. Rather, as claimed by (Asada et al., 2009), such structures should reflect the robot's own process of understanding through interactions with the environment. The authors point out the importance of the social context where the robot is situated in. In fact, the design principles proposed by (Asada et al., 2009) concern also environmental design, that is with how to set up the environment so that the robots embedded therein can gradually adapt themselves to more complex tasks in more dynamic situations (including instructions from a human or a robot) (Asada et al., 2009). In a way, this resembles the scaffolding support that caregivers employ to guide infants' development.

In (Weng et al., 2001), the authors proposed the Autonomous Developmental Robotics paradigm for building developmental robots which consists in four steps: (1) design a body according to the robot's ecological working conditions (e.g., on land or under water); (2) design a developmental program; (3) the robot starts the developmental program (*birth*); (4) humans mentally raise the developmental robot by interacting with it in real time. Autonomous mental development relegates the human to the role of teaching and supporting the robot through reinforcement signals. Moreover, the requirements for a truly mental development include being non-task-specific, because the task is generally unknown at the design time (Weng et al., 2001).

Scassellati proposed an architecture based on an *embodied theory of mind* (inspired by the proposals of (Baron-Cohen, 1995) and (Leslie, 1994)) for the implementation of social learning capabilities in artificial agents (Scassellati, 2001). In particular, he implemented some of the fundamental skills for socially-oriented interaction, including: a visual attention mechanism which combines low-level feature detectors (such as color saturation, motion, and skin color filters) with high-level motivational influences to select regions of

interest; a module for determining whether an observed object is animate or inanimate based on a set of naive physical laws that operate solely on the spatial and temporal properties of the object's movement; a sensorimotor system for detecting faces and for determining the orientation of the person's head, used in directing the robot's attention to the same object that the interacting person is considering.

Sheldon investigated the emergence of communication in artificial agents as an integrated part of a more general developmental progression (Sheldon, 2012). In particular, two main aspects of communication have been addressed, early gestural communication in the form of pointing and spoken language. A developmental progression is implemented with the help of a framework for schema learning. Initially, the robot performs some basic motor babbling and learns the results of its most basic movements, after which they introduced objects for the robot to interact with. Sheldon also investigates the development of proto-imperative pointing gestures and the development of communicative capacity by introducing spoken language.

(Dautenhahn and Billard, 1999) adopted a Vygotskyan approach by assuming that, in order to study the cognitive development of robots, they have to be considered as existing in society. In fact, supporters of situated cognition pointed out that artificial agents, similarly to humans, are not only physically situated but also *socially* situated (Lindblom and Ziemke, 2002). In being situated in a social context, robots are potentially able to acquire knowledge from skilled individuals through social interaction.

The first attempts of teaching robots through social interactions, for instance by demonstrating executions of manual tasks, date back to the 1980s in the field of manufacturing robots. The idea was that robot controllers or behaviours, instead of being manually programmed, can be derived from observing or from being taught by human demonstrators. At the beginning, the interest was motivated by the cost reductions involved by avoiding manual re-programming of robots in factories. Several studies focused on techniques for robot programming by demonstration, also referred as *imitation learning* in the field of robotics (Billard et al., 2008).

The ability to copy motor actions from other individuals is fundamental for the social transmission of knowledge in natural agents. Studies on neonatal imitation, such as (Meltzoff and Moore, 1977), (Meltzoff and Moore, 1997) and (Reissland, 1988), demonstrated that imitation is not a purely cognitive process, as suggested by Piaget, that appears around the age of 12 months. Neonatal imitation is defined as facial, hand, and finger movements and vocalizations made by the newborn in a laboratory environment shortly after an experimenter has demonstrated the same behaviour to the infant (Nagy and Molnar, 2004). However, the authors note, imitation requires the infant to show orientation, attention, learning, effort and motivation when reproducing the previously modelled movements or sounds. Results of the study presented in (Nagy and Molnar, 2004) show that infants are not only capable of responding to a model movement by imitating, but that they also have the capacity to provoke an imitative response, thus sustaining an interaction.

In robotics, efforts still need to be focused on the implementation of such prerequisites for social interaction capabilities in artificial agents. Taking inspiration from previous works, this thesis adopts a developmental approach in providing robots with artificial

mechanisms for the acquisition of motor and basic cognitive capabilities and for the implementation of basic attention skills. In fact, the robotics experiments presented in the following sections are always introduced by a brief review of relevant human developmental studies. Figure 3.1 highlights the robot capabilities that will be presented along this chapter. As briefly discussed in the introduction of this thesis, in Figure 3.1 the skills are elicited along two developmental timelines (*sensorimotor* and *social*) and ordered by complexity. The origin of the two lines indicates the most basic behaviour, namely body babbling (the *zero point*, **ZP**), that has been implemented in the robot as a mean for the acquisition of most of the other more complex motor skills depicted in Figure 3.1.

Section 3.1 reviews developmental and robotics studies investigating exploration behaviours in humans and robots. In addition, it introduces an experiment on motor babbling performed on the humanoid robot Aldebaran Nao for the exploration of the arm action space. Self-exploration mechanisms are adopted also in other experiments presented in the following chapters of this thesis. Section 3.2 illustrates how the humanoid robot Aldebaran Nao equipped with self-exploration behaviours can learn its body maps (**SM1**). Here, body maps are intended as representations of the body, or integrations of different proprioceptive and motor signals (Maravita and Iriki, 2004). As reported also in (Maravita et al., 2003), evidences from animal and human studies suggest that the primate brain constructs various body-part-centred representations of space, based on the integration of visual, tactile and proprioceptive information. Such a constantly updated status of the body shape and posture is fundamental for acting efficiently.

This chapter is also concerned with the topic of joint attention. As defined in (Kaplan and Hafner, 2004), attention is the process whereby an agent concentrates on some features of the environment to the (relative) exclusion of others. In developmental psychology, several studies demonstrated that the development of skills to understand, manipulate and coordinate attentional behaviour lays the foundation of imitation learning and social cognition Tomasello (1995). Joint attention, the capability to share the attention between individuals, is a fundamental skill in social interaction. Unfortunately, it is still an open challenge in the robotics community. As (Kaplan and Hafner, 2004) pointed out, to reach joint attention an agent must understand, monitor and direct the intentions underlying the attentional behaviour of the other agent. The authors identified a number of underlying skills behind the development of joint attention capabilities in robots.

In line with this study, this thesis investigates some of the prerequisites of joint attention, as identified by (Kaplan and Hafner, 2004). As depicted in Figure 3.1 on the social developmental line, two prerequisites for joint attention will be presented in this chapter: the development of imperative pointing gesture skills as a first step towards attention manipulation capabilities and mechanisms for detecting visually salient events such as movements, faces or objects. In particular, Section 3.3 shows how the humanoid robot Aldebaran Nao can exhibit proto-imperative pointing gestures (**SO1**) as a result of failed grasping actions (**SM1**). Section 3.4 shows an implementation of visual saliency detection skills and short-memory system based on a robot ego-sphere (**SO2**). As it will be discussed in Appendix B, the adoption of attention mechanisms, such as visual saliency detection and attention manipulation, can improve human-robot interaction.

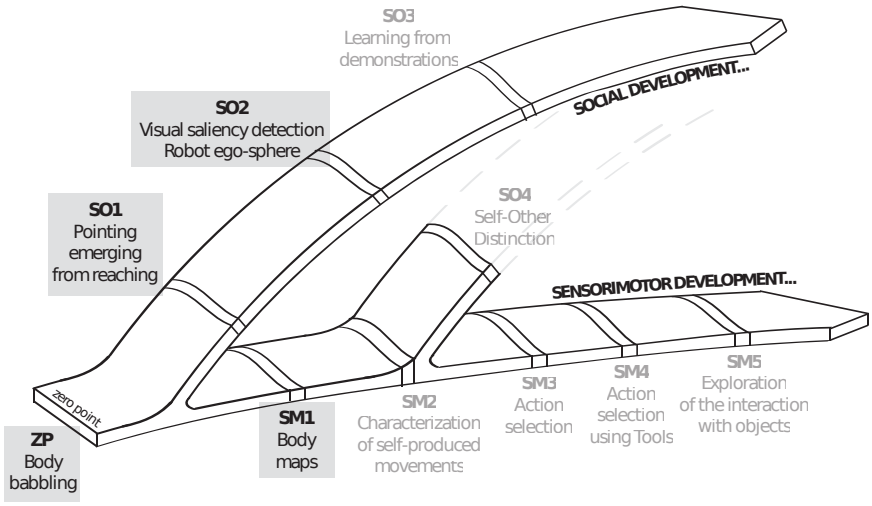


Figure 3.1: This chapter introduces the robot capabilities highlighted in the figure. In particular, Section 3.1 presents mechanisms for self-exploration, namely body babbling (the Zero Point (**ZP**) of the sensorimotor and social developmental timelines), in the humanoid robot Aldebaran Nao. Section 3.2 illustrates how exploration behaviours allow for the building of body maps (**SM1**). Section 3.3 illustrates the development of imperative pointing gestures (**SO1**) out of the reaching capability resulting from (**SM1**). Thus, an implementation of mechanisms for visual saliency detection and of a short-memory system based on a robot ego-sphere (**SO2**) is presented in Section 3.4.

3.1 Body babbling

Infants seem to not have an innate knowledge about what muscle activations achieve a particular perceptual consequence. It is evident that this capability is learned through an experiential process, which (Meltzoff and Moore, 1997) defined as body babbling. During body babbling, infants play with muscle movements which are then mapped to the resulting sensory consequences. In (Kuhl and Meltzoff, 1996), the authors studied this behaviour in the domain of language acquisition. Before the 5th-6th month of age, the vocal tract and the neuromusculature is still immature for the production of any recognizable words and sounds. Through the exploratory behaviour of vocal babbling the infants learn articulatory-auditory relations.

(Jansen et al., 2004) argued that the arise of new skills in infants can be analysed in terms of two developmental parameters: a social dimension and an intentional dimension. From both points of view, babbling falls at the zero-point, as it is a behaviour without social and intentional content. The infant behaviour of body babbling inspired several robotics studies. In (Dearden, 2008), exploration behaviours have been implemented in an artificial agent for gathering evidence to form and to test models for its bodily characteristics. In (Demiris and Dearden, 2005) and in (Demiris and Khadhouri, 2006), the authors propose a way for combining knowledge through exploration and knowledge from others, through the creation and use of mirror neuron inspired internal models (see Chapter 4 for more details).

In (Saegusa et al., 2009), the authors consider motor babbling based sensorimotor learning as an effective method to autonomously develop an internal model of the own body and the environment using multiple sensorial modalities. In particular, they defined a confidence function which works as a memory of reliability for state prediction and control. The aim of this function is to store the reliability of learning result for the sensory input, and exploit it for the next data sampling, so that the robot is able to decide its exploration and learning based on its learning interest.

(Olsson et al., 2006) describes a developmental system based on information theory implemented on a real robot that learns a model of its own sensory and actuator apparatus. The robot develops the model of its sensorimotor system by first performing random movements to create an informational map of the sensors. In particular, the robot builds by motor babbling a library of sensorimotor laws which specify how its possible actions affect its sensors. Using these laws the robot can see motion flows in its visual field and then perform a movement which will have a similar effect.

(Baranes and Oudeyer, 2013) propose an intrinsically motivated goal exploration mechanism which allows active learning of inverse models in high-dimensional redundant robots. Inverse models are learned by performing an exploration driven by the active self-generation of high-level goals in the parametrised task space instead of traditional motor babbling specified inside a low-level control space. Active exploration in the task space leverages the redundancy often characterising sensorimotor robotic spaces.

Similarly to the previous works, in this thesis self-exploration mechanisms have been implemented in robots as a mean for acquiring sensorimotor experience. Starting from a low cognitive level where exploration behaviours are still not goal-directed, random

movement strategies have been adopted for allowing artificial agents to explore their motor capabilities. This section presents the implementation of self-exploration behaviours on the robotic platform Aldebaran Nao [2] published in (Schillaci and Hafner, 2011b) and in (Schillaci and Hafner, 2011a). Aldebaran Nao is a humanoid platform whose dimensions resemble those of a young human subject (see Figure 3.2). The adopted version is provided with 21 degrees of freedom and two head cameras for visual input. Exploration behaviours have been used for the acquisition of arm movement capabilities. The learning session consisted in the robot babbling its arms, that is generating random arm movements.

The Aldebaran Nao robot has been provided with a simple behaviour based on sensorimotor coordination which allowed it to look at its own arm movements. In particular, such a behaviour consisted in the robot moving its arm in a random fashion and, in the meanwhile, generating head movements according to the visual perception of its own hand, or end-effector. While the robot moves its arm, it estimates the position of its end-effector by analysing the visual input and it moves the joints of its neck in order to keep the hand in the center of the image.

However, it has to be pointed out that some assumptions have been made when implementing such a sensorimotor coordination capability. Firstly, the hand of the robot has been tagged with a fiducial marker, which has been used for estimating its 2D position in image coordinates and its 3D position relative to the robot torso. Such a fiducial marker is visible in the bottom pictures of Figure 3.2. The C++ ARToolkit library (http://hitl.washington.edu/artoolkit) has been used in detecting and in estimating the position of the marker.

Secondly, by providing the robot with the capability to visually follow its own movements, a developmental question has been eluded. In fact, it is still under debate to what extent newborns are innately aware of their body and capable of following their own movements. Some researchers demonstrated that eye-hand coordination is present in the newborn's skills repertoire. For instance, (von Hofsten, 1982) showed that newborns, when presented with a moving object in front of them, manifested a primitive visually-guided reaching by moving the hand towards the target. Contradicting Piaget's assumptions, this behaviour shows that visual and manual schemes are not independent in their functioning, but they are integrated from birth (Rochat, 1993). However, this thesis does not deal with the question of to what extent sensorimotor coordination behaviours should be developed or should be given to robots, although this is an important topic that will be addressed in future works, as described in Chapter 5.

In the experiments presented in (Schillaci and Hafner, 2011b) and (Schillaci and Hafner, 2011a), the Aldebaran Nao robot has been provided with a visual attention mechanism based on two modules: a module for detecting fiducial markers and a module for detecting movements[3]. The sensorimotor coordination behaviour consists in the following steps: a motor command, that is, a desired angle position, is sent to each joint of the arm (only one arm is babbling, for each learning session); when the hand of the robot, that is, the

[2]The NaoTH framework has been used for controlling and for programming the robot (http://www.naoteamhumboldt.de).

[3]The module for detecting movements has not been used in the experiments presented in Chapter 4.

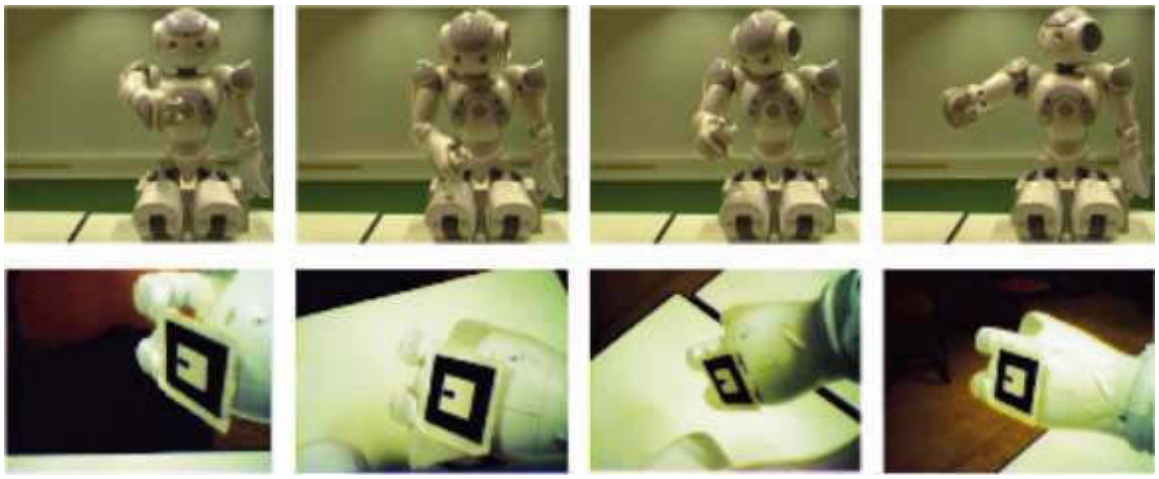


Figure 3.2: A typical babbling sequence from the Aldebaran Nao robot. In the lower part are the corresponding frames grabbed by the onboard camera (note that the camera is placed below the fake eyes of the Nao).

fiducial marker tagging it, is detected, the joints of the neck of the robot are moved in order to center the fiducial marker in the image. During this process, information related to the estimated position of the marker is stored into a knowledge base, together with the current configuration of the arm's and neck's joints. Four joints have been recorded for each arm (shoulder pitch, shoulder roll, elbow yaw and elbow roll) and two joints for the neck (head yaw and head pitch). The bottom camera of the robot have been used for getting visual input.

In the preliminary implementation of the sensorimotor coordination mechanisms presented in (Schillaci and Hafner, 2011b) and (Schillaci and Hafner, 2011a), very often fast arm movements resulted in the fiducial marker going out of the visual field. The motion detection module of the attention system has been implemented for re-catching the sight of moving hand. The motion detection consisted in the following process: while holding the neck joints' positions, the optical flow between sequent image frames is computed[4]. The magnitude of the optical flow is fed into a CAMShift-based tracking function[5]. CAMShift outputs the position of the centroid of the moving area in the camera image. In our case,

[4]Optical flow has been computed using the Lukas-Kanade algorithm from the OpenCV (Open Source Computer Vision) library.

[5]CAMshift (Continuously Adaptive Mean Shift) is a color-based tracking algorithm implemented in the OpenCV library. CAMshift tracks a moving coloured area based on the Mean-Shift algorithm. Usually, a histogram of the colours of the object to track is passed to the function for calculating a probability image, through a histogram back-projection method. Histogram back-projection is an operation that associates the pixel values in the image with the value of the corresponding histogram bin (Allen et al., 2004). Mean-shift is than applied to this image in order to find the centroid of the moving region. As described in (Allen et al., 2004), Mean-shift is a robust, non-parametric algorithm that climbs the gradient of a probability distribution to find the mode (peak) of the distribution. In our case, instead of using histogram back-projection for computing the probability image, the magnitude image resulting from the application of the optical flow algorithm has been fed into the CAMShift algorithm.

Figure 3.3: The mobile manipulator developed at the Space Application Service (http://www.spaceapplications.com/) in the context of the EU-FP7 Marie Curie ITN INteractive RObotics (INTRO). Motor babbling has been used for collecting sensorimotor experience in the mobile robot (Kozlov et al., 2013).

such a moving area in the image is the visible part of the robot's moving arm.

In (Schillaci and Hafner, 2011b) and (Schillaci and Hafner, 2011a), we analysed three random movement strategies (Purely Random (PR), Random Walk (RW) and Inertial Random Walk (IRW)) and experimentally tested on a humanoid robot how they affect the learning speed and how much energy they consume. The results of this analysis is described in Appendix A.

A similar sensorimotor coordination behaviour (based on the Random Walk strategy) has been implemented into the robotic mobile manipulator (see Figure 3.3) developed at the Space Application Service (http://www.spaceapplications.com/) in Belgium in the context of the EU-FP7 Marie Curie ITN INteractive RObotics (INTRO). The motor babbling algorithm presented in this thesis has been integrated in (Kozlov et al., 2013) as part of the integration work for a Urban Search and Rescue (USAR) scenario proposed

within the INTRO project (Jevtic et al., 2012). Motor babbling has been used for collecting sensorimotor experience from the arm of the mobile robot. As in the Aldebaran Nao robot, sensory and motor data consisted in the arm joints angles position and in the end-effector position estimated through the use of fiducial markers. This allowed for learning the inverse kinematic model of the robot's arm through self-exploration.

3.2 Learning body maps through self-exploration

Through babbling and touching their own body, newborns investigate the rich intermodal redundancies, temporal contingencies, and spatial congruence of self-perception (Rochat, 1998). For instance, when manually touching their own face, newborns experience a perceptual event that uniquely identifies their own body as a differentiated object (Rochat, 1998). Such a perceptual intermodal event (that is, occurring from multiple sensory modalities) is the double touch of the cutaneous surface of the face (or other parts of the body) contacted by the cutaneous surface of the hand. This double-touch event does not occur when the baby is touching an external object.

(Butterworth and Hopkins, 1988) reported that the self-stimulation of facial regions by the newborn itself does not produce any rooting reflex. As discussed in the previous section, rooting responses are normally observed when an external object touches particular facial regions of the newborn, such as cheeks. It is thought to be a primitive behaviour supporting the activity of breastfeeding.

(Rochat, 1998) demonstrated that infants, by 3 months, start to show systematic visual and proprioceptive self-exploration and they become sensitive to spatial invariants specifying the self. For example, when feeling their own legs moving in a particular direction, they expect to *see* their legs moving in a similar direction (Rochat, 1998). The findings reported in (Rochat and Morgan, 1998) suggest that infants, by the first year of age, express a sense of their body as a perceptually organised entity which they monitor and control. In other words, these results can be seen as an early expression of a calibrated intermodal space of the body, or of a perceptually based *body schema* (Rochat, 1998).

In (Schillaci and Hafner, 2011b) and (Schillaci and Hafner, 2011a), a robot has been equipped with motor babbling as the learning strategy for mapping its random arm movements with its head movements, determined by the perception of its own body. The experience that the robot collects through body babbling has been used for calibrating the multimodal space of its body. With the acquired mappings, processes similar to those described in (Rochat, 1998) can be implemented. For example, the robot can estimate the position of its hand from the perception of its arm joints configuration, based on what it has experienced in the past.

In this preliminary experiment, learning the mapping between the proprioceptive sensory data and the visual acquired information consisted in collecting the data through body babbling into a knowledge base. Each element of the knowledge base contained the following information: $[hand - position; position - of - neck's - joints; position - of - arm's - joints]$, where the position of the hand is estimated using fiducial markers. Knowing its own body map can be a powerful tool for the robot. For example, given a

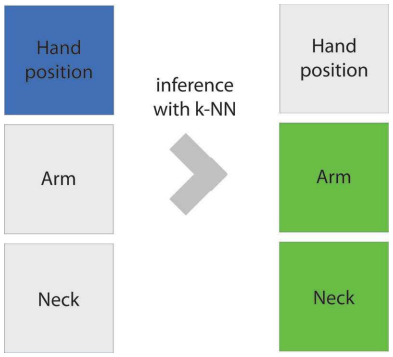


Figure 3.4: Illustration of an inverse prediction. The input of the prediction process is a desired hand position. The prediction process outputs the configuration of the arm and neck joints which result in the end-effector as close as possible to desired position and in centering the marker in the image. The quality of the prediction depends on the experience gathered by the robot during motor babbling.

desired point in the hand's action space, a learned body map can be used to predict the neck's and arm's configurations which let the visually detected marker (representing the hand) be as close as possible to the desired point.

In (Schillaci and Hafner, 2011a), a mapping between the proprioceptive data, represented by the 6D vector $[position-of-neck's-joints; position-of-arm's-joints]$[6], and the visually perceived data, represented by the (x, y) image coordinates of the marker placed on the hand of the robot, has been used to perform a simple forward prediction: given a configuration of the neck and arm joints, infers where the position of the hand will be.

A k-Nearest Neighbours (k-NN) based algorithm has been adopted as inference tool for the forward predictions. In particular, k-NN has been used for performing searches in the knowledge base gathered during motor babbling. OpenCV FLANN library has been used for implementing the k-NN search. Inverse and forward predictions can be performed through k-NN based search on the body maps. For example, as depicted in Figure 3.4, an inverse prediction consist in determining the arm's and neck's joints configuration that results in the hand placed in a desired position (in image coordinates). On the other side, a forward prediction consists in estimating the resulting hand position of a given arm joints configuration.

The inference mechanism is based on a k-NN search in the knowledge base collected during body babbling. For instance, predicting the arm joints configuration which results in the hand placed in a desired position with using a 5-NN consists in the following steps (see Figure 3.7). The desired hand position is given as input to the algorithm. A 5-Nearest Neighbours based search finds the 5 closest vectors in the knowledge base. For each vector, the elements related to the arm joints configuration are extracted. The

[6] 2 Degrees-of-Freedoms for the neck (head yaw and head pitch) and 4 Degrees-of-Freedoms for the arm (shoulder pitch, shoulder roll, elbow yaw, elbow roll).

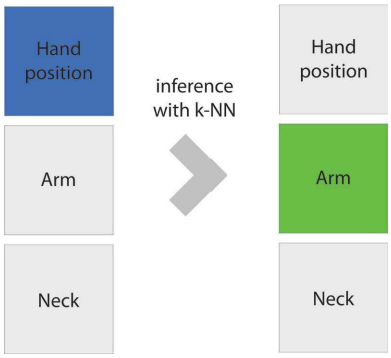


Figure 3.5: Illustration of a simpler inverse prediction. The input of the prediction process is a desired hand position. The prediction process outputs the configuration of the arm joints which result in the end-effector as close as possible to desired position.

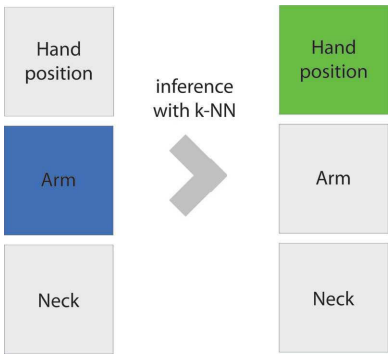


Figure 3.6: Illustration of a forward prediction. The input of the prediction process is a desired configuration of the arm joints. The prediction process outputs the end-effector position resulting from the execution of motor command specified as the input of the process.

outcome is predicted as the mean of these values. Figure 3.8 illustrates the process in inverse prediction with a general k-NN.

Preliminary results on the prediction performance have been collected from babbling samples using the Purely Random (PR), Random Walk (RW) and Inertial Random Walk (IRW) movement strategies. Appendix A illustrates the details. Performance has been evaluated by measuring the prediction errors of forward predictions. In particular, test samples have been extracted from the knowledge base collected during babbling. For each sample, the stored arm joints configuration has been sent to the motors of the robot. The resulting hand position has been compared with the predicted hand position of a forward prediction whose input was the actual arm joints configuration. Figure 3.9 illustrates the process.

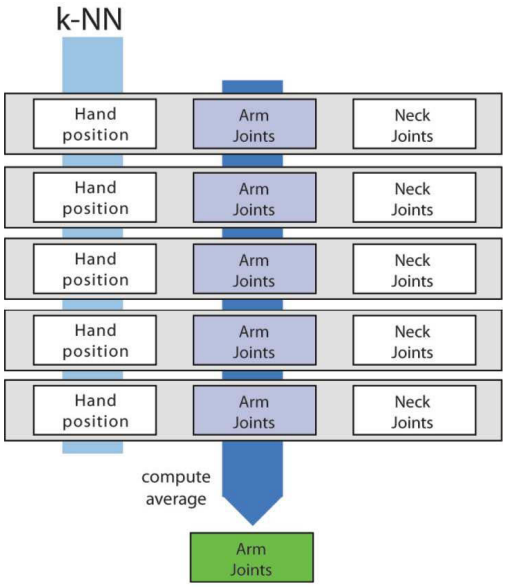


Figure 3.7: Example of a 5-NN search for an inverse prediction. The desired hand position is given as input to the algorithm. A 5-Nearest Neighbours based search finds the 5 closest vectors in the knowledge base. For each vector, the elements related to the arm joints configuration are extracted. The outcome is predicted as the mean of these values.

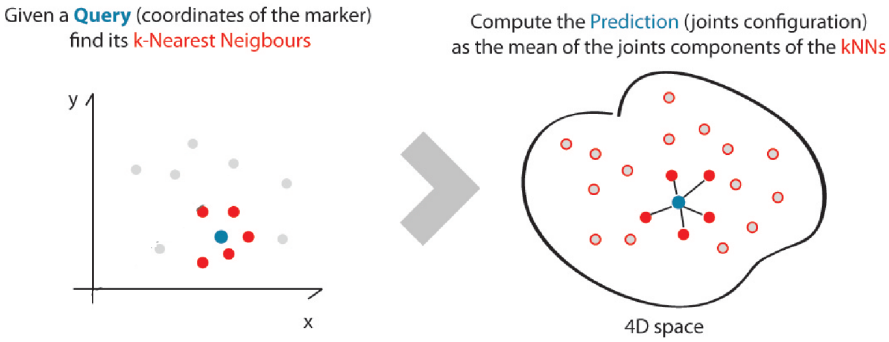


Figure 3.8: Illustration of an inverse prediction. The left graph illustrates the x-y space of the hand positions. Points represent hand positions collected during motor babbling. Each point has corresponding arm and neck joints configuration. The right graph illustrates the 4D space of the arm joints. Points are joints configurations recorded during babbling. The input of the inverse prediction process is a desired hand position (blue point on the left graph). A k-NN search finds the closest elements in the x-y space (red points on the left graph). The corresponding points in the arm joints space are selected (red points on the right graph) and the average of their positions is computed (blue point on the right graph).

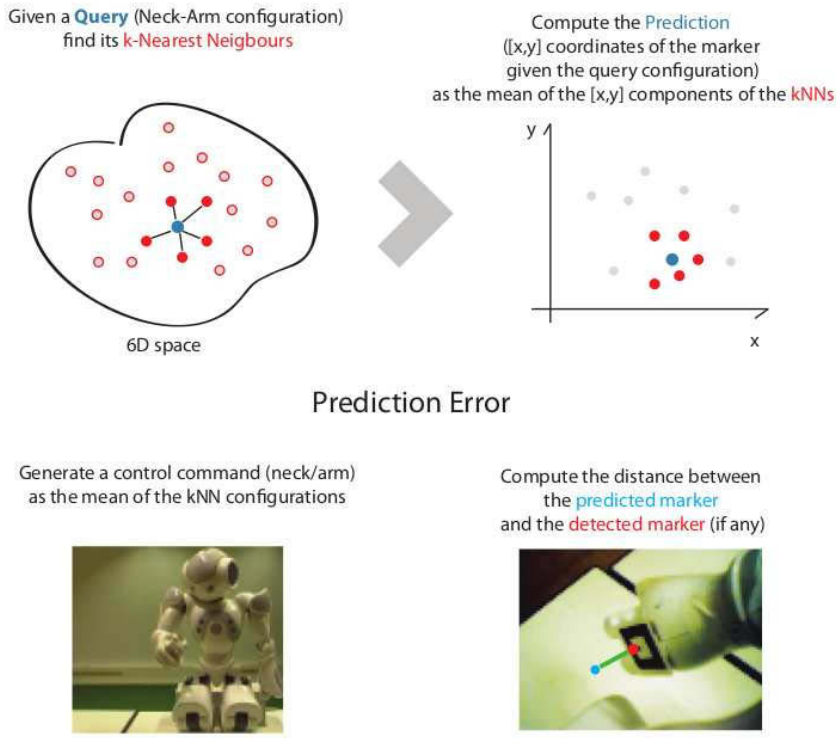


Figure 3.9: An illustration of the calculation of the error of a forward prediction. Performance has been evaluated by measuring the prediction errors of forward predictions. In particular, test samples have been extracted from the knowledge base collected during babbling. For each sample, the stored arm joints configuration has been sent to the motors of the robot. The resulting hand position has been compared with the predicted hand position of a forward prediction whose input was the actual arm joints configuration.

3.3 Is pointing emerging from grasping?

The Russian psychologist Lev Vygotsky sustained that, during the early months of development, biological factors in infants start to be shaped by the social context (Lindblom and Ziemke, 2002). According to his theory of cognitive development, the external social environment has a strong importance in the development of higher mental functions. Also Piaget sustained that development is supported by the external world, but higher mental functions are the result of an internal development occasionally supported by external actors. On the other side, Vygotsky believed that higher mental functions emerge from interpersonal processes after which they are internalised by the child (Lindblom and Ziemke, 2002).

An example of such internalization process is the development of pointing gestures. At the beginning, the pointing gesture seems to be a failed attempt of a grasping movement towards an object. Caregivers often react to such a behaviour by giving the object to the

child. After prolonged repetitions of these reaching attempt/caregiver reaction, the failed grasping behaviour seems to acquire, for the child, the meaning of a pointing gesture. Thus, Vygotsky suggests, it starts to represent an interpersonal connection between the child and the caregiver.

Pointing behaviour is an important skill that develops during early infancy. The capability to perform pointing gestures and to recognise them is a fundamental tool for sharing the target of attention between individuals. As discussed in the previous sections, joint attention is a prerequisite for social interaction (see (Tomasello et al., 2005) and (Kaplan and Hafner, 2004) for a review on joint attention studies). However, it is still under debate how *imperative* pointing gestures, that is, pointing gestures that do not carry any meaning apart from the goal of reaching an object, become meaningful (or *declarative*) gestures. In fact, it is not clear whether declarative pointing gestures origin from imperative pointing ones or not. It starts with imperative pointing around the age of 9 months (Baron-Cohen, 1995). This form of pointing seems to share the same goal of a reaching actions: obtaining a desired object. It is performed by the child regardless of whether an adult is present in the room or is actually looking at the child. Imperative pointing turns into declarative pointing around the age of 12 months, when the child starts to use it to draw somebody else's attention towards an object of interest.

Since imperative pointing is not directly used to draw attention, it is possible that it arose from grasping objects within the reach of the child and turned into pointing for objects that were outside the field of grasp. It is interesting to note that a pointing gesture by someone else is not understood by the child until about the age of 18 months (Butterworth, 1995). Some studies showed that there is no relation between producing pointing gestures and understanding them (Desrochers et al., 1995). This hints to the conclusion that the two skills develop independently from each other and strengthens the hypothesis that pointing may arise from grasping and is not learned by imitation.

In order to test this hypothesis, an experiment has been set where a humanoid robot learns to reach objects by building a body map during random body babbling (Hafner and Schillaci, 2011). This section reports such an experiment. As described in the previous section, self-exploration mechanisms have been used by the humanoid robot Aldebaran Nao for learning the mapping between different sensory modalities. Such a mapping has been used in equipping the robot with predicting abilities of sensory consequences (the position of the hand of the robot) from control commands applied to its neck and its arm (Schillaci and Hafner, 2011a). In a second phase, we equipped the robot with prediction capabilities of arm movement commands that allowed for and resulted in pointing towards an object presented outside the reach of the robot.

In the reaching experiment described above, the body maps are composed by multimodal information taken from 4 degrees of freedom of the arm of the Aldebaran Nao robot, 2 degrees of freedom of the neck and from the visual input from the camera. An external object has been tagged with a fiducial marker. As described in the previous section, during the motor babbling phase, the marker is attached to the hand of the robot and a mapping between the point in joint space and the marker position in the image (2D phase) or in the world (3D, egocentric) is learned. The robot performs random arm movements and maps its arm joints configuration with the position of its end-effector, estimated by analysing

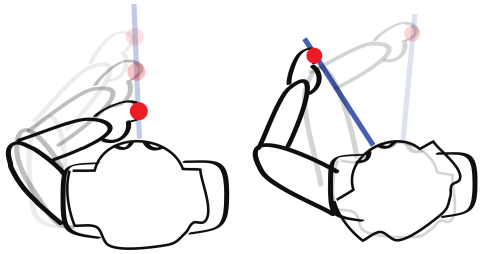


Figure 3.10: Different configurations can result in the same perceived marker position.

the frames grabbed from its head camera. In (Hafner and Schillaci, 2011), a Random Walk babbling strategy has been used (see Appendix A). An inferential tool based on a k-Nearest Neighbour algorithm with $k = 5$ has been used for estimating arm joints configuration for reaching target positions with the hand.

During early exploration behaviour, babies bring objects close to the face and use different sensory modalities for discovering the novelty of the object (such as touching it, biting it, looking at it). Proprioception seems to play a bigger role in hand-eye coordination for infants than for adults (McCarty et al., 2001). In such a developmental phase, it seems that distance is irrelevant as infant vision first needs to develop. We tried to reproduce this inability in our first experiment presented in (Hafner and Schillaci, 2011), where the robot learned the mapping using only the image position of the marker. In this phase, the robot is not able to generate any pointing gestures: an infinite number of arm and neck configurations could result in the same perceived marker position in the image (see Fig. 3.10). Depth perception seems to start around the age of 5 months in human infants. This new ability has been simulated by letting the robot learn to use the estimated 3D marker positions with respect to the torso coordinate frame.

Babbling sessions have been performed in two experimental setups: on the real robot and on the Webots robot simulator (similarly to the real set-up, the robot's hand has been tagged with a marker). In the experiments presented in (Hafner and Schillaci, 2011), a 30 minutes babbling session running in the Webots simulator resulted in 14068 collected data points. A 20 minutes babbling session on the real robot resulted in 7531 collected samples. Both knowledge bases have been used for generating pointing gestures.

In the testing phase, objects tagged with AR markers have been presented outside the field of reach of the robot, but within its field of view in a distance of up to 1 meter from the robot's head. The implemented robot behaviour consists of predicting and generating the arm joints configuration that results in the shortest distance between the hand and the detected marker. In this phase, the hand of the robot is not tagged with a fiducial marker. The arm joints configuration can be estimated by searching the k nearest neighbours (in this case, 5-NN) to the estimated object position in the babbling knowledge base. When the target object resides out of the field of reach of the robot, the nearest neighbours are found on the hull of a sphere representing the robot's field of reach (see Figure 3.11). Thus, the implemented algorithm outputs an arm joints configuration that results in placing the hand as close as possible to the object, in fact resembling a pointing gesture (see Figure

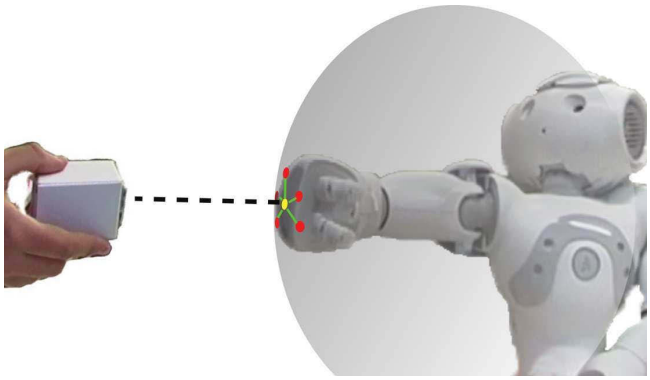


Figure 3.11: An illustration of how pointing gestures are generated. k-NNs are expected to be on the hull of the field of reach (here represented as an ellipsoid).

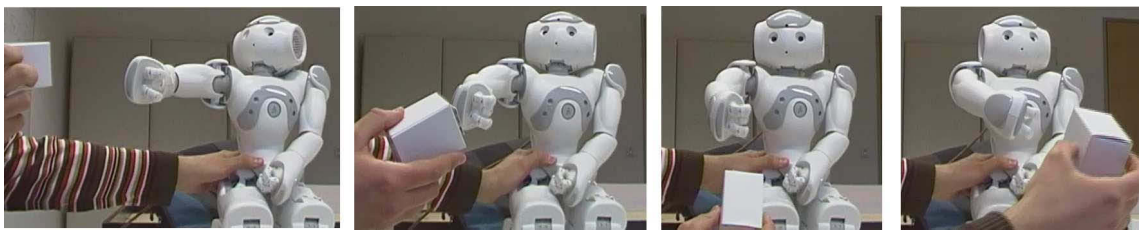


Figure 3.12: A sequence of pointing gestures. An AR marker is placed on the non-visible surface of the object held by the user. An interesting side effect of the body babbling is that the robot automatically follows with its gaze when trying to reach an object.

3.12).

"How can pointing emerge from grasping behavior (T2.3)?" has been one of the open issues in developmental psychology and developmental robotics identified by Kaplan and Hafner (Kaplan and Hafner, 2004). (Hafner and Schillaci, 2011) present an approach to identify the necessary sensory and informational prerequisites for realising this effect using body babbling together with a simple prediction model. In the experiments, the robot's behaviour that was based on learning to reach for an object automatically resulted in imperative pointing behaviour when the object was outside the reach of the robot.

3.4 Visual saliency detection

Human-robot interaction often fails due to the fact that the robot and the human try to communicate about different things and the human partner has wrong expectations of his/her robotic partner. Current social robotic systems require interaction protocols which decrease the intuitiveness of the interaction itself, causing frustration and despair in the user. Several prerequisites have been identified (see (Kaplan and Hafner, 2004) and (Schillaci and Hafner, 2011a) about the features (both physical and cognitive) that let a robot interact effectively and naturally with a human user. (Schillaci et al., 2013a) stresses

the fact that robots need to reach joint attention with the users for having successful interactions. This has not been achieved so far, since joint attention not only requires visual attention on the same visual features in the environment, but also skills in attention detection, attention manipulation, social interaction skills and even intentional understanding (Kaplan and Hafner, 2004).

Attention is a cognitive skill, studied in humans and observed in some animal species, which lets a subject concentrate on a particular aspect of the environment without the interference of the surrounding. There is evidence from developmental psychology studies that the development of skills to understand, manipulate and coordinate attentional behaviour lays the foundation of imitation learning and social cognition (Tomasello, 1995).

This section reports the study carried out in (Bodiroža et al., 2011) and (Schillaci et al., 2013a), where two prerequisites for joint attention in a humanoid robots have been presented, namely attention manipulation and visual saliency detection. In the previous section, an implementation of the capability to generate imperative pointing gestures has been presented. This section illustrates how, by creating a saliency based attentional model and combining it with a robot ego-sphere, a robot can engage in an interaction with a human and starting an interaction game including objects.

In our world, we are constantly surrounded with items, such as objects, people and events, which stand out to their neighbouring items. This is represented with the saliency of those items. Saliency detection represents an attentional mechanism, through which those items are discovered, and it enables humans to shift their limited attentional resources to those objects that stand out the most (Bodiroža et al., 2011). The ability to orientate rapidly towards salient visual events has evolutionary significance because it allows the organism to detect quickly possible prey, mates or predators in the visual world (Itti and Koch, 2001).

As discussed in (Bodiroža et al., 2011), two main approaches for the development of mechanisms for saliency detection and visual attention in artificial agents can be found in the literature. One approach, proposed by (Itti and Koch, 2001), relies on saliency form of focal bottom-up attention. In this method, the visual input is analysed using pre-attentive computations of visual features. In biological vision, pre-attentional mechanisms for computing visual features are found throughout many visual areas and the visual thalamus, such as retina, superior colliculus, lateral geniculate nucleus and early visual cortical areas (Suder and Wörgötter, 2000). Neurons at the earliest stages are tuned to simple visual attributes such as intensity, contrast, orientation, direction and velocity of motion (Itti and Koch, 2001). Thus, it seems that the incoming visual input is decomposed in the early stages of visual processing by a set of feature-selective filtering processes.

In (Itti et al., 1998) and (Itti and Koch, 2001), a model of a visual attention system has been presented, where visual input is first decomposed into a set of topographic feature maps, or saliency maps. Different spatial locations then compete for saliency within each map and only the locations that stand out can persist (Itti et al., 1998). In the bottom-up process, all the saliency maps feed a master saliency map over the entire visual scene.

The second approach is based on a top-down process influenced by motivation. While bottom-up detection uses different low-level features (e.g. motion, colour, orientation and intensity) for saliency detection, top-down detection relies on high-level features, and it

is highly influenced by our current goals and intentions. The combination of bottom-up and top-down processes is highly inspired by similar mechanisms in humans (see (Itti and Koch, 2001) and (Treisman, 1985)). Moreover, the ability to distinguish between animate and inanimate objects, to recognise if a moving object is a person to interact with, to detect and follow eye gaze are essential prerequisites in the development of joint attention and of social interaction skills (Baron-Cohen, 1995).

In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), we implemented an attention mechanism based on visual saliency detection[7]. Saliency detection represents a process of image analysis during which regions of interest are detected in the visual scene. The visual saliency detection mechanism consisted in a set of feature detectors, or filters, to be applied to the visual input. Each filter generates a saliency map from the visual input. A combination of the outcoming saliency maps is calculated, in order to let the robot direct its attention to the point which has the highest saliency. Three filters have been implemented in (Bodiroža et al., 2011) and (Schillaci et al., 2013a): face detection, motion detection and object detection.

Infants show visual preference towards human faces. Some studies lead to the conclusion that infants are born with some information about the structure of faces (Morton and Johnson, 1991). (Walton et al., 1992) showed that newborns of 12-36 hours of age produced significantly more sucking responses in order to see an image of their mothers' faces as opposed to an image of strangers' faces using a preferential sucking procedure. Although it is still not clear what is the nature of such a preference, researchers seem to converge to the claim that face-like stimuli recruit in newborns more visual attention that other salient events (Meltzoff and Moore, 1997). In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), a face detection filter has been implemented for the robot's visual saliency detection mechanism. The OpenCV implementation of the Viola-Jones algorithm has been adopted for detecting and tracking faces.

Some developmental studies, as for example (Moore et al., 1997), suggest that motion information, such as head turn movements, facilitates infants' learning of joint attention. Infants are also attracted by moving objects. In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), a motion detection filter has been implemented in the visual saliency detection mechanism using a Farneback's optical flow algorithm. As for the face detection filter, the OpenCV library has been adopted for computing the optical flow from subsequent images. In the experiments presented in (Bodiroža et al., 2011) and (Schillaci et al., 2013a), fiducial markers have been used for tracking objects. An object detection filter has been implemented in the visual saliency mechanism for detecting fiducial markers.

In robotics, visual saliency detection mechanisms based on saliency maps and on a multi-modal salient ego-sphere have been proposed, such as in (Peters et al., 2001) and (Fleming et al., 2006). A multi-modal salient ego-sphere is a tessellated sphere, where each of the nodes (vertex on the surface of the tessellated sphere) represents one point in

[7]In (Bodiroža et al., 2011), attention manipulation mechanisms have been integrated with visual saliency detection mechanisms. As it will be described in the following, a tessellated ego-sphere has been implemented by Saša Bodiroža (Early Stage Researcher fellow of the EU-FP7 Marie Curie INTRO ITN, at the Humboltd-Universität zu Berlin) where salient events can be projected on, thus implementing a short-memory mechanism for the robot.

the surrounding space of the robot. Input images are analysed by using saliency detection filters, such as the aforementioned ones. The mean value of salient regions from the resulting maps are then assigned to the nodes, which are closest to the real position of the projection of the salient region to the sphere surface. The ego-sphere enables a robot to shift its attention from one salient area to another one in an apparently random and natural fashion.

(Ruesch et al., 2008) describes a framework based on saliency detection for the humanoid robot iCub[8] with adopting a robot ego-sphere. However, instead of using a tessellated ego-sphere, the authors adopt a matrix projection of the ego-sphere, which leads to higher precision of the perception of the world. However, this also increases the computational complexity, due to required image transformations and a higher number of arithmetic operations required per iteration.

Similarly to the previous works, in (Bodiroža et al., 2011) and (Schillaci et al., 2013a) an ego-sphere has been adopted as a short-term memory system for storing the salient events detected by the robot from its surrounding environment. The sphere is centred at the robot's neck coordinate system, while the saliency map of its surrounding is projected onto the sphere's surface [9]. Through mechanisms of habituation, inhibition and forgetting of salient areas the robot is able to explore its surroundings, and by finding areas of maximum saliency, it locates the next area to be attended. Furthermore, the robot uses pointing to the currently attended object, as a first step towards joint attention (Kaplan and Hafner, 2004).

The approach adopted for implementing the robot ego-sphere is similar to that of (Fleming et al., 2006). Projection of events on the surface of the ego-sphere and search space are reduced by tessellating the sphere and storing information about salient areas in the edges of the tessellated sphere. Figure 3.13 shows an illustration of the ego-sphere around the robot's body, as a short-term memory system where salient events are stored in the edges of the tessellated sphere.

However, such an approach introduces errors in projection of salient areas, because projection space is reduced to a set of nodes on the sphere surface. As described in (Bodiroža et al., 2011), the mean saliency of a salient area is computed and assigned to the closest point, which is found by using the nearest neighbour search. Figure 3.14 illustrates the projection of a salient event onto the ego-sphere. The closest node to the projection of the salient event position is selected for updating.

While the matrix projection approach adopted by (Ruesch et al., 2008) has higher precision and finer representation of salient areas, the approach adopted in (Bodiroža et al., 2011) and (Schillaci et al., 2013a) is faster, due to the lower number of arithmetic operations that need to be performed during the projection and the search.

As explained in (Bodiroža et al., 2011), the sphere tessellation is performed by recursive division of triangle faces of an icosahedron. By increasing the recursion depth, which

[8] For more information about iCub, see http://www.icub.org/

[9] The information coming from face and motion detection filters are stored in the ego-sphere, keeping track from which of those two channels information come from. In the implementation presented in (Bodiroža et al., 2011) and (Schillaci et al., 2013a), objects' positions are estimated using fiducial markers and are not stored in the ego-sphere, rather in a separate list of elements.

Figure 3.13: An illustration of a tessellated ego-sphere for the Aldebaran Nao robot.

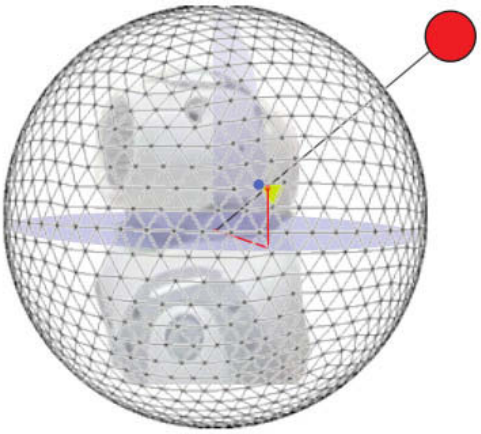


Figure 3.14: Illustration of the projection of a salient event onto the ego-sphere. The closest node to the projection of the salient event position is selected for updating.

represents the number of recursive calls to the tessellation function, each initial face is divided into a higher number of smaller faces, achieving a higher number of nodes and smaller error of projection of salient nodes. The relation between the recursion depth and the tessellation frequency, used in (Peters et al., 2001), is $f_t = {d_r}^2$, where f_t is the tessellation frequency and d_r is the recursion depth. The tessellation frequency represents the number of edges which connect centers of two neighbouring pentagons. The implementation uses a recursion depth of 4, resulting in 2562 nodes, with the mean theoretical

projection error being 1.15° and maximum 2.65°. In addition to the recursion depth, other three parameters characterise the tessellated ego-sphere: the habituation and inhibition weights and a decay step for the forgetting process.

As described in (Bodiroža et al., 2011), habituation, inhibition and forgetting mechanisms are employed in order to favour shift of attention to information in new locations. This is inspired by the inhibition of return mechanism in spatial attention in humans. (Posner et al., 1985) demonstrated that humans, when generating saccades, tend to inhibit orienting toward visual locations which have been previously attended (inhibition of return). [10]

Habituation is the process during which the robot gets used to the attended point, which results in loss of interest in that point. In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), it is modelled with the following function:

$$h(t) = h(t-1) + w_h\dot{(}1 - h(t-1)), \tag{3.1}$$

where $w_h \in [0, 1]$ represents the habituation weight.

When habituation to a certain salient point exceeds a predefined habituation threshold level, inhibition is turned on with a change from 1 to 0, which results in the appearance of a shift of attention to the next most salient point. The reported saliency of a point is the product of its current inhibition value and its saliency.

However, inhibition has a temporary nature and it is updated with an inhibition weight, according to the following function:

$$i(t) = i(t-1) + w_i\dot{(}1 - i(t-1)), \tag{3.2}$$

where $w_i \in [0, 1]$ represents the inhibition weight.

[10] In (Bodiroža et al., 2011), the parameters for the algorithm were chosen so that the overall ratio of accuracy and speed is optimised. The camera resolution is set to 320×240 pixels. However, the image is resized for different filters. In the case of face detection in a typical experimental setup, there is enough level of detail after scaling down by 2. The image for motion detection is scaled down by 4, which is inspired by the human peripheral vision. Our peripheral vision is blurred, but it provides good results for motion detection.

We adopted some parameters tuned in order to avoid tracking motion caused by noise. Dense Optical Flow has been computed using the Farnebäck algorithm, which outputs an image with the same size as the original image containing the magnitude of the movement for each pixel of the image. First, we set a minimum flow magnitude threshold in order to prevent adding points to the ego-sphere generated by noise. We tuned empirically the value to 7 (velocity in pixel). Then, we extracted blobs from the flow magnitude image, taking into account only blobs bigger than a given size (50 pixel, in a 80×60 image), in order to cut out isolated moving pixels or small areas probably generated by noise.

Minimum face size for the Viola-Jones algorithm is set to 35×35 pixels and the scale factor is set to 1.4 (faces detected up to $1.5m$).

Habituation and inhibition weights, habituation threshold level, decay step for forgetting influence for how long the robot will focus on one salient area and how fast it will forget older salient areas. Such parameters have been chosen so that the robot seems responsive enough in a typical interaction. The habituation weight, w_h, is lower for the face detection compared to the motion detection. For the inhibition weight, w_i, the relation is reversed. This results in a behaviour where the robot habituates faster to motion, since it is present for a short amount of time at one location, while the face usually does not move much in the usual direct interaction. In the current implementation, weights are $w_{h,face} = 0.4$, $w_{h,motion} = 0.7$, $w_{i,face} = 0.5$, $w_{i,motion} = 0.15$, and the decay factor is $d = 0.2$.

Habituation and inhibition weights affect the speed of the respective processes. For certain filters, such as motion, locations of salient regions will quickly change in time. Motion can be used to initially draw attention, but it should also decay faster, because it is only present for a short period of time. Other filters can be employed to detect different salient objects in and around these areas, such as face or color detection. Under the assumption that the face does not move much during the interaction, a robot should have slower habituation to locations that contain faces.

In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), the values of salient locations decay according to the decay factor $d \in [0, 1]$, which enables a form of short-term memory for the robot. After each iteration, each salient location will have a lower value than before, thus newer information will have higher saliency than older information. A partially preprogrammed motivation system has been implemented to show how different behaviours can result in the activation or deactivation of parts of the attention system, actually implementing a top-down approach for saliency detection, or in the activation of attention manipulation.

Pointing is a way for manipulating the attention of someone else. As discussed in the previous section, recognising and performing pointing gestures is very important for being able to share attention with another person (Kaplan and Hafner, 2004). In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), attention manipulation mechanisms through pointing gestures have been implemented using the techniques described in the previous section.

In (Bodiroža et al., 2011) and (Schillaci et al., 2013a), we implemented a partially preprogrammed motivation system by which the robot can change its behaviour due to its current beliefs and desires. In addition, the experiments presented in (Schillaci et al., 2013a) aimed at several goals: test the quality of the implemented saliency detection and attention manipulation mechanisms; identify those physical and behavioural characteristics that need to be emphasised when implementing attentive mechanisms in robots; measure the user experience when interacting with a robot equipped with attentive mechanisms; find correlations between heterogeneous robot features perceived by the participants during the exhibition of attentive mechanisms; and analyse the differences in the perception depending on the different behaviours performed by the robot.

Several metrics for measuring Human-Robot Interaction can be found in the literature, from measuring the ability of a robot to engage in temporally structured behavioural interactions with humans (Jonsson and Thorisson, 2010), to evaluating robot social effectiveness from different points of view (engineering, psychological, sociological) (Steinfeld et al., 2006). In (Schillaci et al., 2013a), we adopted a series of metrics based on cognitive science studies about measuring social skills in humans and based on studies about how robots are perceived by humans and whether this perception affects the expectation humans have about robot intelligence (Godspeed questionnaire (Bartneck et al., 2009)).

Quantifying human behaviour usually requires the analysis of video recordings, questionnaires and interviews. In (Schillaci et al., 2013a), we used the first two methods for quantifying the quality of robot behaviour. We set up four interaction experiments between a humanoid robot and a user and recorded them. After each experiment, the user was asked to fill a questionnaire on the quality of the interaction and on the perception of

several functional and physical properties of the robot.

(Takayama and Pantofaru, 2009) adopted part of the Godspeed questionnaire in their measurements, finding that people who held more negative attitudes toward robots felt less safe when interacting with them. They also studied human personal space around robots, finding that experience with owning pets decreases the personal space that people maintain around robots, experience with robots decreases the personal space that people maintain around robots, and a robot looking at people in the face influences proxemic behaviours. The latter suggests to perform proxemics analysis when measuring attentive mechanisms in robots.

The implementations presented in (Schillaci et al., 2013a) have been tested in four combinations of activated parts of the attention system, which resulted in four different behaviours:

A. **Exploration.** In this motivation state, all the saliency filters are activated, the ego-sphere is updated frame by frame, and control commands are applied to the joints of the head to let the robot focus on salient events. As depicted before, the robot is attracted by movements, faces and objects.

B. **Interaction.** In the interaction phase, the robot is not exploring anymore using face and motion filters. Its behaviour is now focused on looking and pointing at the object (such movements are generated using the predictive model explained before).

C. **Interaction avoidance.** In this state, the robot just detects markers and, if any, moves its head toward a configuration far from the current one, actually trying to look towards areas which do not contain any markers. This behaviour has been implemented for trying to convince the user that the robot is bored and it does not want anymore to follow the interaction session.

D. **Full interaction.** This behaviour is composed of a sequence of the previous behaviours. The first performed action is *exploration*. Once the robot has detected a person to interact with and an object which can be used to draw the attention of the user, its motivation state changes to *interaction*. We set an interest variable which decreases over time and which specifies the lapse of time the robot stays in the interaction state. We noticed in a previous interaction experiment that users used to bring the object and hand it to the robot for the whole session. Using this interest decay variable, we can stop the interaction phase and change the robot's behaviour to *interaction avoidance*. The interest variable (initialised as 1) decrease slowly (by 0.005) when the person is handing the object to the robot or rapidly (by 0.025) if not. We estimate that the user is handing the object to the robot if the marker position is moving. When the interest factor goes below zero, we programmed the robot to change its behaviour to *interaction avoidance*. After a while, the current behaviour is set back to *exploration*. When the interest factor goes below zero, we programmed the robot to change its behaviour to *interaction avoidance*. After a while, the current behaviour is set back to *exploration*.

Figure 3.15: Experimental setup showing interaction between the Nao and a person.

The experiments consisted of the robot performing the behaviours described before in four separate interaction sessions, one per each of the four behaviours. The experiment supervisor manually activated or deactivated them. Figure 3.15 shows a frame taken from a typical interaction session. The user sat in front of the robot at a distance of ca. 90 *cm*. For each person, each interaction test lasted one minute. We recorded the interaction with a standard camera (resolution 640×480) placed at ca. 2 meters perpendicularly to the robot-user axis. Beside the table where the robot was standing there was a scale drawn on a whiteboard for the visual estimation (estimated average error: $5cm$) of the distance between the nose of the user and the head of the robot and from the hand of the user and the head of the robot; according to the type of interaction, we noticed that the users move their hands closer to the robot. After each of the four interaction sessions, the participants were asked to fill a questionnaire about the quality of the interaction with the robot and about the perception of robot behaviours.

Appendix B shows the details of the statistical analysis performed on the collected data. In synthesis, the robot's level of interactiveness has been found to be positively correlated with user experience factors like excitement and robot factors like lifelikeness and intelligence, suggesting that robots must give as much feedbacks as possible in order to increase the intuitiveness of the interaction, even when performing only attentive behaviours. This was confirmed also by proxemics analysis: participants reacted more frenetically when the interaction was perceived as less satisfying. Improving the robot's feedback capability could increase user satisfaction and decrease the probability of unexpected or incomprehensible user movements. Finally, multi-modal interaction (through arm and head movements) increased the level of interactiveness perceived by participants. Positive correlation has been found between the elegance of robot movements and user satisfaction.

Chapter 4

Sensorimotor learning and simulation of experience

Chapter 3 addressed the first research question of this thesis: what are the basic behavioural components an artificial agent should be provided with for being able to develop motor and cognitive capabilities?

This chapter addresses the second and third research questions that have been pointed out in the introduction of this thesis:

- How can an artificial agent represent and store the experience generated through the basic behaviours identified in the previous chapter?

- How can the acquired experience be reused and what computational processes are needed for generating basic cognitive skills out of it?

In particular, the aim of this chapter is to identify a framework for representing and storing experience in artificial agents, to test such a framework in different experimental sets and to use it for generating basic cognitive capabilities.

Chapter 2 of this thesis introduced two paradigms. Firstly, for understanding and implementing intelligence in artificial agents, it is necessary to study them in their relation with the environment. Theories on embodied cognition consider behaviours and cognition as processes that emerge from a strict coupling between agent and environment. The way how an agent interacts with its environment is strongly influenced by its bodily characteristics (Pfeifer and Bongard, 2006). It has been discussed that defining models of robots' embodiment and their surrounding world *a priori* should be avoided, since the risk is to stumble across problems such as robot behaviours lacking adaptability and the capability to react to unexpected circumstances. Thus, Chapter 3 investigated mechanisms for autonomous acquisition of sensorimotor experience in artificial agents. In particular, exploration behaviours based on sensorimotor coordination have been implemented for allowing a robot to gather experience about its corporeality. In answering the second research question of this thesis, the established framework of internal models has been adopted (Wolpert and Kawato, 1998). In robotics, such a framework, which allows for

coding *sensorimotor* experience in artificial agents, has not been fully studied and exploited. This thesis contributes to a deeper investigation of internal models by making the least assumptions possible while constructing them. In particular, inverse and forward models have been trained with different set of low-level sensory and motor data generated by the robot through exploration behaviours, or demonstrated by a human, or acquired through kinesthetic teaching.

The second paradigm introduced in Chapter 2 is that mental simulations of sensorimotor experience could serve as computational mechanisms for the implementation of cognition in robots. The studies reviewed in Section 4.1 suggest that simulation processes could be behind the functioning of basic cognitive processes in humans. In particular, modal simulations, such as recreations of perceptual, motor and introspective states, could account for the off-line characteristics of cognition, in which internal simulations of sensorimotor cycles are executed (Barsalou, 2008). The adoption of the internal models framework is motivated by its capability to generate simulations of sensorimotor cycles. As it will be shown in the following experiments, internal simulations have been used as computational processes behind the implementation of basic cognitive capabilities in a humanoid robot.

The rest of this chapter is structured as follows. Section 4.1 reviews studies supporting the existence of simulation mechanisms in human cognitive processes. Section 4.2 introduces the framework of the internal models and it describes how internal simulations can be performed. In addition, it presents a review on the existing computational models and implementations of simulation mechanisms in robotics, with or without the explicit adoption of the internal models framework. Section 4.3 presents the experiments that have been carried out during my Ph.D. studies that share the following mechanisms: sensorimotor learning; coding of sensorimotor experience through internal models; simulation processes for implementing basic cognitive capabilities. Figure 4.1 lists the basic cognitive capabilities that have been implemented in the humanoid robot Aldebaran Nao using sensorimotor simulation mechanisms. In particular, Section 4.3.1 shows an implementation of action selection capabilities, in which the robot is capable of selecting one of the two arms to use for reaching a desired target point, based on its past sensorimotor experience and on simulating the outcome of its motor actions. Section 4.3.2 present an extension of the previous experiment, in which the robot's morphology is extended with the use of a tool. Section 4.3.4 presents an account on self-other distinction capabilities based on mental simulations of hand trajectories, which are characterised in Section 4.3.3. Section 4.3.5 introduces experiments where the robot learns basic interactions with objects and stores the gathered sensorimotor experience into an internal models framework. Two learning paradigms are tested, as described in Section 4.3.5: self-exploration or learning from demonstrations. In particular, an experiment on the recognition of motor behaviours involving interaction with objects is presented, where such behaviours are learnt through self-exploration. Section 4.3.5 concludes with presenting a similar experiment, where the same motor behaviours are learnt through observing human demonstrations. As described in this last experiment, internal simulations can be used also for detecting the target object of a motor action.

The work presented in this chapter has been published in (Schillaci et al., 2012a), (Schillaci et al., 2012b), (Schillaci et al., 2013b) and (Schillaci et al., 2013c).

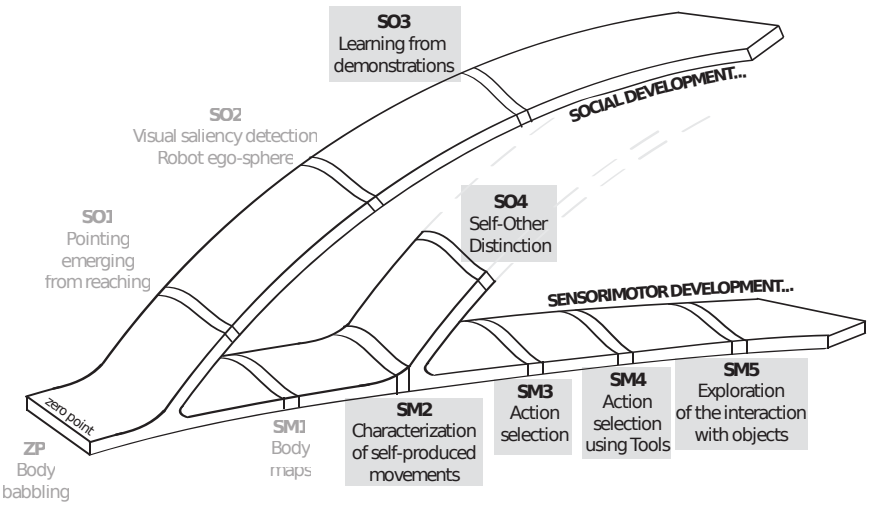


Figure 4.1: This chapter introduces the robot capabilities highlighted in the figure. Section 4.3.1 shows an implementation of action selection capabilities (SM3). Section 4.3.2 presents an extension of the previous experiment, in which the robot's morphology is extended with the use of a tool (SM4). Section 4.3.4 presents an account on self-other distinction capabilities (SO4) based on mental simulations of hand trajectories, which are characterised in Section 4.3.3 (SM2). Section 4.3.5 shows an experiment on the recognition of motor behaviours involving interaction with objects, where such behaviours are learnt through self-exploration (SM5). In adddition, a similar experiment is presented where the same motor behaviours are learnt through observing human demonstrations (SO3).

4.1 Studies on sensorimotor simulations

In the last decades, several studies suggested that mammalian brains implement sensorimotor prediction processes (Wexler and Klam, 2001). For instance, (O'Keefe and Recce, 1993) found that particular cells in the hippocampus of the rat's brain, which are thought to be involved in the representation of the animal's position, fire in a way that is not constantly correlated with the phase of the sinusoidal EEG theta pattern. The first burst of firing consistently occurs at a particular phase of the reference theta, but each successive firing burst moves to a point earlier in the theta cycle, as the rat runs through the field (O'Keefe and Recce, 1993). This suggested that the current position of the animal is periodically anticipated along the path.

(Eskandar and Assad, 1999) studied the role of monkey's posterior parietal cortex in the visual guidance of movements. Monkeys were trained to use a joystick to guide a spot to a target. Visual and motor influences were dissociated by transiently occluding the spot and by varying the relationship between the direction of joystick and spot movements. In the lateral intraparietal area of the monkey's brain, the authors found cells which were not selectively modulated by either visual input or motor output, but rather seemed to encode the predicted visual trajectory of the occluded target.

Human studies also suggested that sensorimotor prediction processes exist in motor planning and execution. (Wolpert et al., 1995) asked participants to move their arm in the absence of visual feedback, in order to test whether the central nervous system is able to maintain an estimate of the hand position. Each participant gripped a tool that was used to measure the position of the thumb and to apply forces to the hand using torque motors. The hand was constrained to move along a straight line. Each participant was shown the initial hand position. Then, the light was turned off and the participant was asked to move the hand either to the left or right. During the movements, random assistive or resistive force field was generated by the torque motor. Once the movement terminated, the participant was asked to indicate the visual estimate of the unseen thumb position using a trackball held in the other hand. The distance between the actual and visual estimate of thumb location was recorded as a measure of the state estimation error. The bias of this location estimate, plotted as a function of movement duration, showed a consistent overestimation of the distance moved (Wolpert et al., 1995). The temporal dynamic of the bias systematically showed an increase of the error during the first second of movement and, then, a decay. The authors proposed that the initial phase is the result of a predictive process that estimates the hand position, followed by a correction of the estimate when the proprioceptive feedback is available.

(Spivey et al., 2000) demonstrated that oculomotor responses can be triggered even in absence of any visual stimulus. The authors showed that participants constructing mental models of complex visual scenes (when no visual information is available) tend to make the same eye movements that would be made when viewing a particular scene.

(Tourville et al., 2008) presented a functional magnetic resonance imaging (fMRI) study investigating the neural mechanisms underlying auditory feedback control of speech. Participants were asked to speak monosyllabic words under two conditions: normal auditory feedback of their speech and perturbed auditory feedback condition, in which the first

formant frequency of their speech was unexpectedly shifted in real time. The authors found compensations to the shift in the acoustic measurements and a greater neural activity in posterior superior temporal cortex bilaterally during the perturbed feedback, suggesting that such a region could code mismatches between expected and actual auditory signals.

(Wexler and Klam, 2001) asked participants to estimate displacements of an occluded moving target, where the movement was driven by the observer's manual action, or passively observed. The authors found that when the observer actively caused the target to move by a manual rotation of a knob, predictions were farther advanced (or more anticipatory) than in passive trials. Decreasing the congruence between motor action (e.g. rotations) and visual feedback (e.g. translations) diminished, but did not eliminate, the anticipatory effect of action.

In (Frak et al., 2001), subjects were requested to determine the feasibility of grasping an object placed at different orientations. The authors demonstrated that the opposition axis (defined by the final finger position through which opposite forces operate on the object) is affected by limb biomechanics and by the visual characteristics of the object itself. In addition, they found that the time to give the response is a function of the object's orientation, suggesting that the subject must mentally move his arm in an appropriate position before the response can be given (Jeannerod, 2001).

Sensorimotor predictions are thought to be involved in higher cognitive skills than just motor control. In (Tucker and Ellis, 1998), participants had to decide as fast as possible whether an object displayed on a screen was upright or inverted. The authors found that left-right orientation of common graspable objects had a significant effect on the speed with which a particular hand made a simple push button response, even though the horizontal object orientation was irrelevant to response determination. The orientations of the objects were chosen so as to make them preferentially compatible with a reach-and-grasp movement by the left or right hand.

(Wexler et al., 1998) performed a study on mental rotation where participants were shown two visual stimuli that differed by a rotation and possibly by a reflection. The subject was asked to decide whether one pattern was a reflected version of the other. A solution strategy would consist in rotating the pattern with the hand or the head and seeing the result. However, in mental rotation tasks, the action is planned but not executed and the perceptual result of the planned rotation is simulated (Wexler et al., 1998). The authors found that motor rotations (using a joystick handle) made a concurrent mental rotation of a visual image faster and more accurate if the two rotations were in the same direction, slower and more error prone if they were in opposite directions. When the two rotations were in the same direction, faster motor rotation tended to speed up the mental rotation, while slower motor rotation tended to slow it down (Wexler and Klam, 2001). This suggests that the motor system and predictive mechanisms are involved in mental rotations.

(Bosbach et al., 2005) demonstrated that two subjects lacking cutaneous touch and sense of movement and position show a selective deficit in interpreting another person's anticipation of weight when seeing him/her lifting boxes. The authors suggested that this ability occurs through mental simulation of action dependent on internal motor representations, which require peripheral sensation for their maintenance.

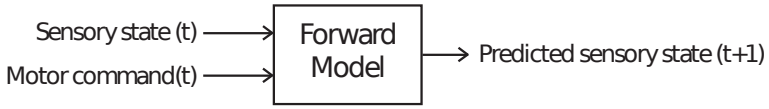


Figure 4.2: An illustration of the forward model (predictor).

(Jeannerod, 2001) proposed that the motor system is part of a simulation network that is activated under a variety of conditions in relation to action, either self-intended or observed from other individuals. The function of this process of simulation would be not only to shape the motor system in anticipation to execution, but also to provide the self with information on the feasibility and the meaning of potential actions.

A faulty functioning of self-monitoring mechanisms is thought to be responsible for some of the symptoms present in schizophrenia (Frith, 1992). The authors highlight the importance of predictive mechanisms for giving agents a sense of agency and more importantly, of the changes in the external world due to the agents' actions.

The studies reviewed before suggest that simulation processes could account for off-line characteristics of cognitive processes. Indeed, the capability of recreating perceptual, motor and introspective states could improve robot's interactive skills. The next section reviews studies on computational internal models, that could provide simulation mechanisms in robots.

4.2 Existing implementations

In the literature on motor control in primates and in humans, the theoretical concept of internal models has been proposed (Wolpert et al., 1995). An internal model is a system that mimics the behaviour of a natural process (Miall and Wolpert, 1996). Two main types of internal models have been proposed: forward and inverse models. A forward model is an internal model which incorporates knowledge about sensory changes produced by self-generated actions of an agent. For example, in the kinematic domain, a forward model represents the mapping between the arm joints position and the endpoint coordinates of the hand (as described in the previous chapter). In general, such a model would aim to represent the normal behaviour of the motor system in response to motor commands (Miall and Wolpert, 1996). In particular, given as input a sensory state S_t that the system is perceiving and a motor command M_t (an intended or actual action), the forward model outputs, or *predicts*, the next sensory situation S_{t+1} that the system will perceive after the actuation of the motor command (see Figure 4.2). Forward models were first proposed in the control literature as means to overcome problems such as the delay of feedback on standard control strategies and the presence of noise, both also characteristic of natural systems (Jordan and Rumelhart, 1992).

While forward models (or predictors) present the causal relation between actions and their consequences, inverse models (or controllers) perform the opposite transformation providing a system with the necessary motor command M_t to go from a current sensory situation S_t to a desired one S_{t+1} (see Figure 4.3). In fact, motor control involves trans-

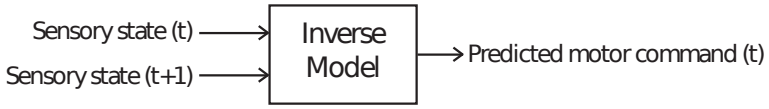


Figure 4.3: An illustration of the inverse model (controller).

formations, or translations of behavioural objectives into muscles activations (Atkeson, 1989). Inverse models encapsulate knowledge about the behaviour of the motor system (Miall and Wolpert, 1996), that is the knowledge of the causal events, in terms of motor commands, that produce particular states transitions. For example, in the kinematic domain, an inverse kinematic model outputs the set of arm joint angles that achieve a desired hand position.

As described in the previous section, similar processes have been found to be capable of modelling several behaviours and characteristics of the brain. Evidences also suggest the locations of brain areas involved in these processes. For example, (Blakemore et al., 1998b) suggested that the cerebellum implements simulation mechanisms. In particular, the authors used fMRI to examine neural responses when subjects experienced a tactile stimulus that was either self-produced or externally produced. The authors found more activity in somatosensory cortex when the stimulus was externally produced. In the cerebellum, less activity was associated with a movement that generated a tactile stimulus than with a movement that did not. The authors suggested that this difference is due to the involvement of the cerebellum in predictive mechanisms of the sensory consequence of the movements. When a movement is self-produced, the estimate of its sensory consequences can be better predicted and this prediction can be used to attenuate the sensory effect of the movement (Blakemore et al., 2000). (Goodbody et al., 1998) suggested that sensory input and motor output signals are combined to provide an internal estimate of the state of both the world and one's own body, and that such an estimate may be stored in posterior parietal cortex. In fact, the authors reported that a patient with a lesion of the superior parietal lobe showed both sensory and motor deficits consistent with an inability to maintain such an internal representation between updates.

In robotics, internal models can play an important role in the development of cognition, as they naturally fuse multiple sensory modalities together with motor information providing agents with multi-modal representations (Wilson and Knoblich, 2005). Moreover, internal models can provide an artificial agent with a tool for performing simulations of sensorimotor cycles. As discussed previously, mental simulations of sensorimotor experience could be the basis of some of the off-line characteristics of cognition, as suggested by theorists of grounded cognition (Barsalou, 2008). The joint and coordinate action of both forward and inverse models as depicted in Fig. 4.4 gives an agent a practical sense of situations and can even account for subjective experience as a ground for consciousness (Kiverstein, 2007).

Much research has been done on computational internal models for action preparation and movement. Efforts have been focused on hand-arm trajectories, where these models present an elegant solution to the problems posed by systems that have to deal with dif-

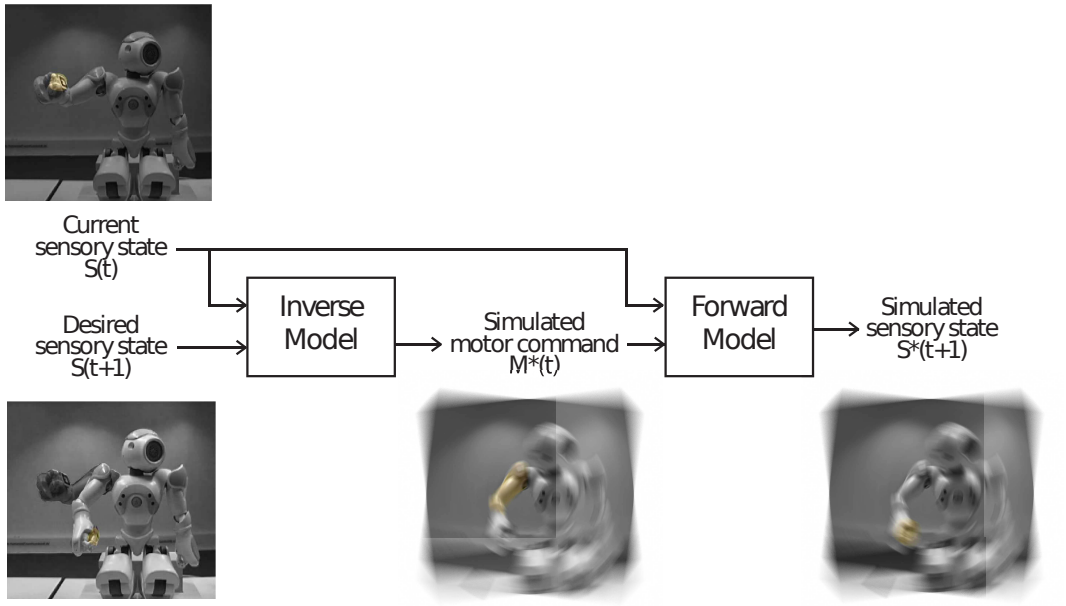


Figure 4.4: Example of an internal simulation. The inverse model simulates the motor command (in the example, a displacement of the joints of its arm) needed for reaching a desired sensory state, from the current state of the system. Before being sent to the actuators, such a simulated motor command can be fed into the forward model which anticipates its outcome, in terms of sensory perception. A prediction error of the internal simulation can be calculated by comparing the simulated sensory outcome $S*_{t+1}$ with the desired sensory state S_{t+1}.

ferent contexts (Wolpert and Kawato, 1998), such as the handling of objects with different weights (Wolpert and Ghahramani, 2000). With the use of Hidden-Markov models, the main proposal became a standard reference known as the MOdular Selection And Identification for Control (MOSAIC) model (Haruno et al., 2001). In MOSAIC, different pairs of inverse and forward models provide motor commands according to what they learned. In biological systems, such transformations are often not known a priori. A knowledge such as the one modelled by a forward model is generally learned through a prolonged exploration of the outcomes associated with particular choices of actions (Miall and Wolpert, 1996). In MOSAIC, a responsibility estimator weights the contribution of each pair to choose a motor command according to the context and the behaviour the system is currently modelling. The authors proposed the model as a base for more complex behaviours and actions in hierarchical MOSAIC (Wolpert et al., 2003). HMOSAIC is capable of accounting and model social interaction, action observation and action recognition. The authors tested the architecture in motor learning and control through a simulation task, where a hand had to track a given trajectory while holding three different objects with defined characteristics. The authors assumed the existence of a perfect inverse dynamic model of the arm for the control of reaching movements (Haruno et al., 2001). The controller needed to learn the motor commands to compensate for the dynamics of different objects. In the following sections, a similar internal models based architecture will be presented for providing a humanoid robot with basic cognitive capabilities.

In cognitive robotics, very interesting results have been presented by (Dearden and Demiris., 2005), where a robot learns a forward model that successfully imitates actions presented to its visual system. The task was to learn the forward model for the movement of a robot's grippers. In particular, the robot needed to learn to predict the effects of its motor commands. The experiment consisted in the robot babbling with its motor commands (three available commands: open, close or stop a gripper), observing what happened (velocity of the gripper) and then learning the relationship between the motor command and the observation using a Bayesian network. The forward model was thus used to allow the robot to imitate simple hand gestures of a human. Dearden also presented a more complex system where a robot learns from a social context by means of forward and inverse models using memory-based approaches (Dearden, 2008). In the following sections, similar experiments will be presented, where a humanoid robot acquired sensorimotor experience through self-exploration, learning from demonstration or from being manually moved by a human teacher (kinesthetic teaching). As it will be discussed in the following, both the inverse and forward models have been trained using the self-experienced sensorimotor data. In our experiments, sensorimotor simulations processes have been adopted for implementing basic cognitive capabilities out of the acquired set of inverse-forward models pairs.

(Demiris and Khadhouri, 2006) proposed the HAMMER architecture (Hierarchical Attentive Multiple Models for Execution and Recognition) based on inverse-forward model pairs for action execution and action understanding. HAMMER is coded using Bayesian Belief Networks and has also been extended to include cognitive processes such as attention. Similarly to MOSAIC, HAMMER relies on the concept of motor simulation for implementing mechanisms of action recognition and execution. The recent discovery of

the mirror neuron system in primates (Rizzolatti et al., 1996) and humans (Grezes et al., 2003) suggested that the motor system is involved in the perception of other's actions. In particular, the mirror neuron system would form a shared neural substrate for motor action observation and execution. One of the proposals is that simulation processes are involved in the recognition of demonstrated motor actions. (Demiris and Khadhouri, 2006) implemented a similar mechanisms in HAMMER. A mobile robot platform equipped with an on-board camera observed a human demonstrator performing object-oriented actions. The task of the robot was to match the observed actions with the equivalent in its actions repertoire. Inverse-forward model pairs were coding each action in the repertoire. In particular, a number of inverse models were implemented using predefined primitives provided by the robot manufacturer. Forward models were hand coded for each of the inverse models, using kinematic rules to output a qualitative prediction of the next state of the system for each of these inverse models (Demiris and Khadhouri, 2006). When observing a human demonstration, information such as hand and object position were feeding the inverse models (a selective process has been also implemented). Efferent copies of motor commands were sent to the forward models which provided as output a predicted sensory state. A prediction error was calculated and used to update the confidence of the corresponding inverse model. As it will be described in the next sections, a similar mechanism for action recognition has been implemented in (Schillaci et al., 2012b), where different learning strategies are compared for inverse-forward model pairs. Self-exploration and learning by demonstration mechanisms have been implemented for gathering sensorimotor data to be used in training inverse and forward models pairs. Moreover, predefined motor primitives have not been used in coding inverse models. Rather, controllers have been trained using the experienced sensorimotor data.

(Akgun and Tunaoglu, 2010) presented a computational model capable of recognising actions on-line by modifying the Dynamical Movement Primitives (DMP) framework (Ijspeert et al., 2001) and by using simulation mechanisms. DMPs are non-linear dynamic systems which are used for imitation learning, action generation and recognition. In (Akgun and Tunaoglu, 2010), a robot is trained with demonstrated actions such as approaching an object from different sides. The learning consists of estimating parameters of a function to be used during action generation. Action recognition is performed by processing the observed action as a new one to be learned, estimating the parameters of the function approximator and, thus, comparing them to the ones in the repertoire. However, such an approach requires that the whole action is observed, in order to calculate DMPs' parameters and to use them for recognition (Akgun and Tunaoglu, 2010). The authors modified the original DMP framework to allow online action recognition. For recognition, initial state observations are given to action generation systems as initial conditions and future trajectories are simulated. Errors are calculated by comparing the simulated trajectories with the observed ones, which are then used to compute the recognition signals. Recognition signals could be interpreted as the likelihood of the observed motion to be the corresponding simulated action (Akgun and Tunaoglu, 2010).

Simulation mechanisms have also been proposed in the context of robot navigation. For example, (Hoffmann and Möller, 2004) and (Hoffmann, 2007) presented a chain of forward models that provided a mobile agent with the capability to select different actions

for achieving a goal situation and to perform mental transformations during navigation. In goal planning, usually a sequence of motor commands is required. The authors approached the problem of searching for a motor sequence as an optimization in a chain of identical forward models, implemented as multilayer perceptrons (MLP). In particular, the input to the chain was the current sensory situation. The optimisation problem aimed at finding the sequence of actions, whose last stage matches the desired sensory situation and the free parameters were the motor commands for each stage (Hoffmann and Möller, 2004). A similar model was used to recognise different scenarios in simulation, by means of long term prediction (Möller and Schenck, 2008).

(Lara et al., 2007) adopted a chain of forward models to avoid collisions in a navigation task and to let the robot obtain information on the ownership of its actions. This was achieved using data from a simulated agent to train a forward model and was afterwards tested on a real robot. The visual input in these experiments was the data coming from a simulated linear camera. A major drawback of this approach is the need of three time steps to disassociate size and distance of objects in the field of view (Lara and Rendon, 2006).

In (Ziemke et al., 2005), experiments on internal simulations of perception have been performed on a mobile robot. The authors did not address explicitly their system as exploiting internal models. However, they incorporated several aspects of the sensorimotor theories and implemented internal simulations to achieve a navigation task. In their experiment, collision-free corridor following behaviours have been trained by a simulated Khepera robot. Neural networks have been also used to predict the next time step's sensory input as accurately as possible, in order to let the robot act blindly, i.e. repeatedly using its own prediction instead of the real sensory input.

In the context of mental training, (Di Nuovo et al., 2013) presented a model of a robot controller that allows a humanoid robot to autonomously improve its sensorimotor skills. This was achieved by endowing a neural controller with a secondary neural system. By exploiting the sensorimotor skills already acquired by the robot, the secondary neural system generated additional imaginary examples that could be used by the controller itself to improve the performance through a simulated mental training.

This thesis approaches the problem of implementing basic cognitive capabilities in robots by investigating the use of sensorimotor simulation processes. In particular, experiments on a humanoid robot using internal models and sensorimotor simulation mechanisms will be presented. Sensorimotor simulations have been adopted as a computational process for implementing basic cognitive capabilities in the humanoid robot Aldebaran Nao. The implemented predictive capabilities are a result of the sensorimotor experience that the robot collected in the training session.

4.3 Experiments

This chapter deepens the study on the multi-modal internal models framework (Wolpert and Kawato, 1998) for representing sensorimotor behaviours in artificial agents. This thesis contributes to a deeper investigation of internal models by making the least as-

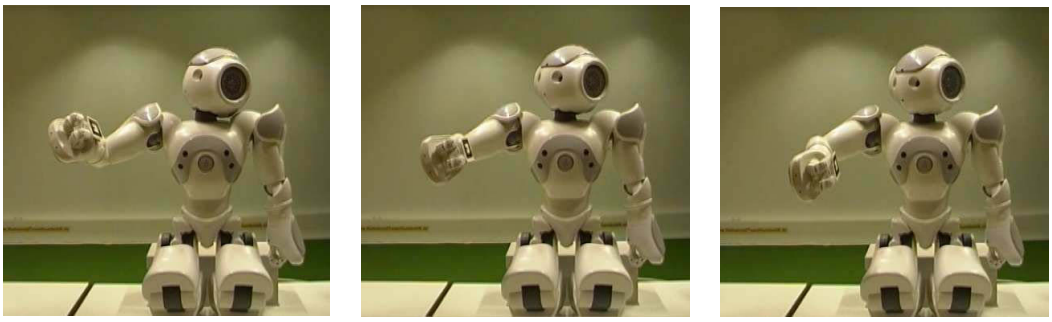


Figure 4.5: Robot during motor babbling

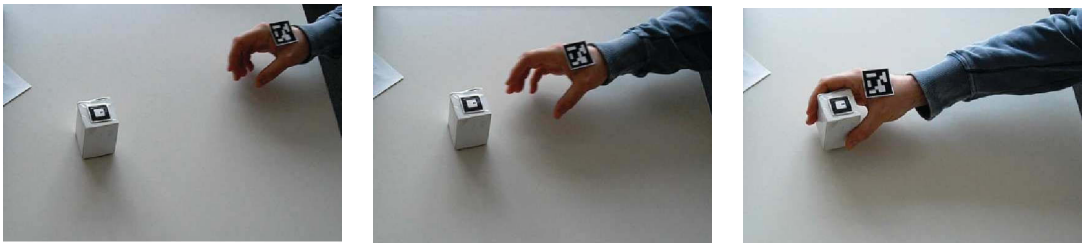


Figure 4.6: Robot observing a skilled human demonstrator

sumptions possible while constructing them. In particular, inverse and forward models have been trained with low-level sensory and motor data generated by the robot through exploration behaviours (Figure 4.5), or with data observed from human demonstration (Figure 4.6) or with sensory and motor data generated by a human physically guiding the robot (kinesthetic teaching, Figure 4.7).

The adoption of the internal models framework is motivated by its capability to generate simulations of sensorimotor loops. As it will be shown in the following sections, internal simulations can be used as computational processes behind the implementation of basic cognitive capabilities in a humanoid robot. I strongly believe that simulation mechanisms have still not been fully studied and exploited in the field of cognitive robotics. Besides the adoption of internal models and the implementation of internal simulation mechanisms, this thesis presents implementations of basic cognitive capabilities in robots following a developmental paradigm (as described in Chapter 3 and as illustrated in Figure 4.1 at the beginning of this chapter).

In the following sections, experiments will be presented that share the following procedures: an artificial agent is equipped with mechanisms for acquiring sensorimotor experience; such experience is coded as inverse(controller)-forward(predictor) pairs; internal simulations mechanisms endow the robot with basic cognitive capabilities based on its past experience.

Figure 4.7: Kinesthetic teaching

4.3.1 Action selection

The first experiment has the robot simply to learn how to move its left and right arms. Out of this knowledge, a very basic skill has been implemented: the capability to choose one of the two arms for achieving a particular task. Here, a simple task has been addressed, that is approaching a point with one of the two hands, or end-effectors. For simplicity, whether to approach it with the left or the right hand depends only on the distance between the target and the hand: the winning strategy corresponds to the arm movement that results in the smallest hand-target distance. Although this is a trivial task in robotics, the aim here is to derive such a skill from the knowledge about arms' movement capabilities that the robot acquired by itself. In particular, this experiment shows that simulations processes can be used in anticipating the outcomes of each of the two movements and, thus, in selecting the most appropriate action. The quality of such anticipations relies on what the robot experienced about its body capabilities.

As described in chapter 3, in Schillaci et al. (2012a) we implemented a motor babbling mechanism for generating sensorimotor experience in the humanoid robot Aldebaran Nao. The robot was programmed to learn its arm motion capabilities for reaching positions in its action space. Learning consisted in producing sequences of random configurations of the arm joints. During the movements, the robot visually estimated the position of its end-effector and it stored the visual information together with the motor commands applied to the arm joints. In the experiment presented here, the robot is programmed to self-explore its action space and to collect information about the hand-joints mapping for each of the two arms. Sensorimotor data gathered from each of the two behaviours is coded as an inverse-forward models pair (see Figure 4.8).

Multi-Layer Perceptrons (MLPs) have been adopted for implementing the internal models[1]. Each model is represented by a separate MLP, which is trained with the data

[1]A Multi-Layer Perceptron is a feedforward artificial neural network that maps a set of input data with a set of output data (Noriega, 2005). An MLP consists of multiple layers of nodes in a directed graph. In this thesis, only 3-layers MLPs have been used, with one input, one hidden and one output layer. Each layer is fully connected to the next one. Each node of the MLP, except the input ones, is a neuron with an activation function. Here, a symmetrical sigmoid activation function has been used. For

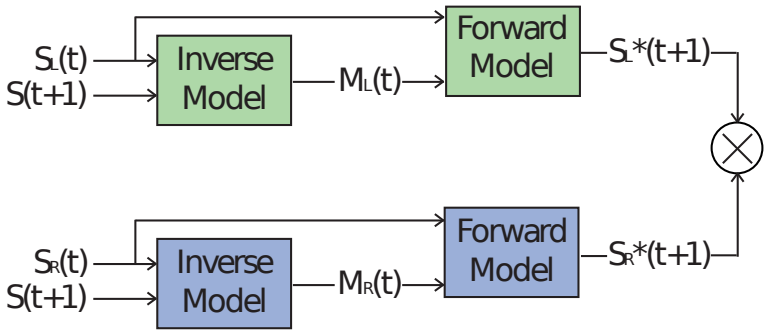


Figure 4.8: Two competitive pairs of inverse and forward models. The light green pair corresponds to the left arm of the robot. The light blue pair corresponds to the right arm of the robot. In this experiment and in the one presented in section 4.3.2, absolute joint positions have been used as motor commands, instead of displacements in the joint space. Thus, the information provided by the sensory state S_t is not really needed by the inverse model for predicting the motor command M_t or by the forward model for anticipating the sensory state S_{t+1}. In fact, in the example of the inverse model, a motor command expressed in terms of *final* arm joints configuration will produce a unique end-effector position, regardless of any initial hand position. For being consistent with the internal models representation, this Figure illustrates S_t as inputs of the models. Nonetheless, it is important to take out from the Figure the flow of the simulation process: each inverse model produces an efferent copy of the motor command, whose effect is simulated by the forward model, which subsequently produces a sensory prediction.

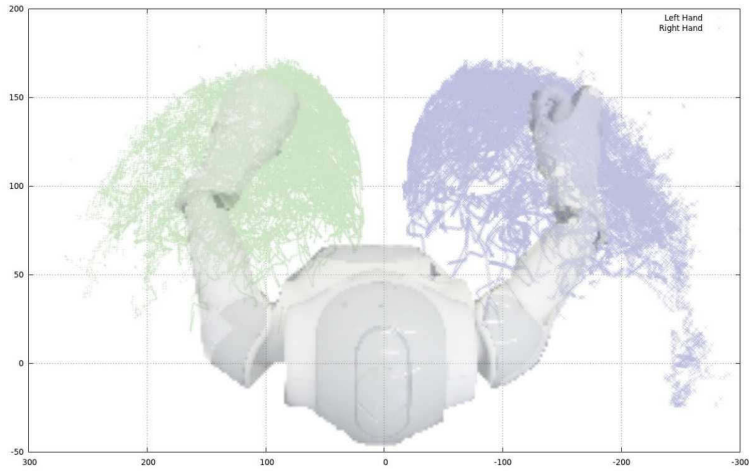


Figure 4.9: Reachable spaces for both hands of the Nao. Each point in the clouds has been experienced together with the motor command that resulted in that end-effector position. Colours identify the arm producing the end-effector position: light green for the left arm, light blue for the right one.

gathered during babbling (for each internal models pair, the same data set is used for training both the inverse and the forward models). In training forward models, sensory and motor data is used as input and sensory data as output. In training inverse models, instead, sensory data is used as input and motor data as output. In this experiment, we used the $[x, y, z]$ coordinates of the robot hand (estimated, as in the previous experiments, through a fiducial marker placed on the end-effector) as sensory state and the positions in angles of the arm's joints (shoulder roll, shoulder pitch, elbow roll, elbow yaw) as motor commands.

Figure 4.9 shows a top-down view of the action spaces of the two robot's arms after a motor babbling session. Each point in the clouds has been experienced together with the motor command that resulted in that end-effector position. Colours identify the arm producing the end-effector position: light green for the left arm, light blue for the right one.

implementing and for training MLPs in the experiments presented in this thesis, the Machine Learning module of the OpenCV library has been used. In particular, a back-propagation algorithm has been adopted as a supervised learning technique.

To compute the network, all the weights of the connections between the nodes need to be known (source: OpenCV Documentation, http://docs.opencv.org/index.html). The training algorithm computes the weights, by taking the training set consisting in input vectors (whose size equals the one of the MLP's input nodes) with the corresponding output vectors (whose size equals the one of the MLP's output nodes) and by adjusting the weights so that the network can output the desired response to the provided input vectors. The sequential back-propagation algorithm has been used as the iterative technique for adjusting the weights according to the training set. As described in the OpenCV documentation, once trained, the MLP can be used for performing predictions, which consists in taking the feature vector as input, passing it as input to the first hidden layer, computing the outputs of the hidden layer using the weights and the activation functions and, thus, passing the outputs further downstream until the output layer is computed.

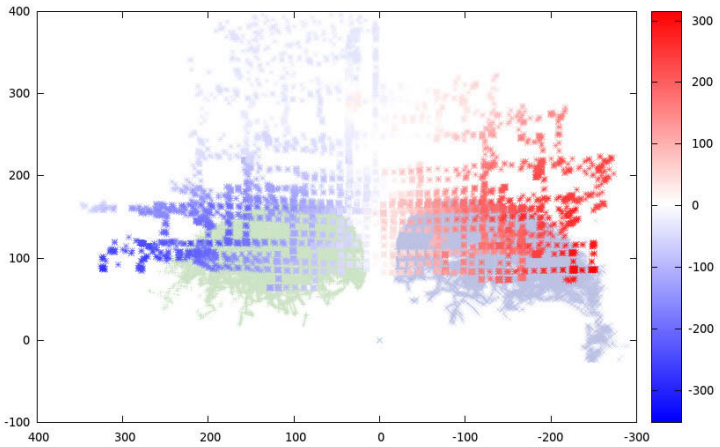


Figure 4.10: Prediction errors of the simulations of the left and right arm movements. Star-like points are the test target positions used for generating internal simulations. Their colours range from dark-blue to red, representing the likelihood that the left arm or the right one has been selected for the execution. White points correspond to target positions where the prediction errors are similar, meaning that both left and right arms could be used for the action execution

As introduced before, a mechanism based on internal simulations of the sensorimotor loop has been implemented for selecting the arm to use when reaching a target point in the space. Simulations of motor commands are performed before executing the actual actions with each arm. By estimating the outcomes of the simulated actions, the robot can select the best movement strategy to adopt. Indeed, the better the robot knows its sensorimotor skills, the more precise are the predictions of the sensory consequences of its actions.

In this action selection experiment, internal simulations are run as follows. The robot is asked to reach a target position with one of its arms. An inference process finds the motor command M_t which brings the system to the desired sensory situation S_{t+1}, composed by the coordinates of the goal position. This process is performed by querying each of the two inverse models (left and right arm) with the desired target position. For each arm, the corresponding simulated motor command is sent to the paired forward model to infer the resulting sensory situation. The two outcomes are the predicted end-effector positions, which are then compared with the target position. Such comparisons produce prediction errors, in this case euclidean distances between the target and the two predicted hand positions. Therefore, the inverse-forward model pair which generates the smallest prediction error is selected for the actual execution and the corresponding motor command which has been predicted by the inverse model is sent to the related arm joints.

Figure 4.10 shows the prediction errors of the simulations of the left and right arm movements in reaching a number of test target points. The green and light-blue clouds are the projections on the horizontal plane of the points collected during babbling (green for the left arm, blue for the right one). They illustrate both arms' action spaces. In (Schillaci et al., 2012b), 73149 babbling samples have been used for training both the inverse and

forward MLPs. For training the right arm MLPs, 56844 babbling points have been used. A back-propagation algorithm has been used in training the Multi-Layer Perceptrons.

In Figure 4.10, star-like points are the test target positions used for generating internal simulations. Their colours range from dark-blue to red, representing the likelihood that the left arm or the right one has been selected for the execution. White points correspond to target positions where the prediction errors are similar, meaning that both left and right arms could be used for the action execution.

4.3.2 Action selection and tool-use

The potentialities of internal simulation mechanisms have been emphasised in a second experiment by changing the morphology of the robot. An extension tool is attached to one of the end-effectors of the robot in order to modify its reachable space. This experiment recalls the study on tool-use performed by (Maravita and Iriki, 2004). Tool use is an important skill that is acquired during early childhood in humans and requires several cognitive abilities related to sensorimotor interaction. (Bril et al., 2010) consider tool-use as an instance of a goal-directed action which requires also a cognition component. However, tool-use is not an exclusive human skill. Several animals are capable of using tools. (Maravita and Iriki, 2004) propose that we hold an adaptive body map comprising body posture and shape. They suggest that the body schema is extended temporarily with the tools we are using.

From this perspective, the extended arm experiment on the robot can be seen as the body of the robot being temporarily extended by a suitable tool for a specific task (namely reaching an object). In (Schillaci et al., 2012a), we expected the emergence of this body map by means of multi-modal sensory representations of the environment coded in the inverse-forward model pairs.

As in the previous experiment, the robot is programmed to acquire the new sensorimotor scheme by performing a new motor babbling session. Figure 4.11 illustrates an approximation of the action spaces for both the arms of the robot, with an extension tool on its left arm. While far-away positions can be reached with the left extended end-effector of the robot, disregarding the side of the robot where they are presented, other positions (such as the ones near the robot's chest) are only reachable with the right arm.

Again, internal simulation processes allow the agent to choose the arm to use in reaching a target position. Figure 4.12 shows the prediction errors of both pairs when presented with the same target positions as in Figure 4.10. Here, each inverse and forward models pair have been trained with $130,151$ babbling points for the extended arm and with $56,844$ babbling points for the right one. The use of the extension tool emphasises the potentials of the internal simulations. The most remarkable examples were presented when the goal position was located on the right half of the workspace but relatively far from the robot. In these situations the best model (the one with the least error) would be the one coding for the arm with the tool (left arm). This behaviour was different from the one reported in the previous experiment, where under the same situation the robot would have executed a reaching action with the right hand.

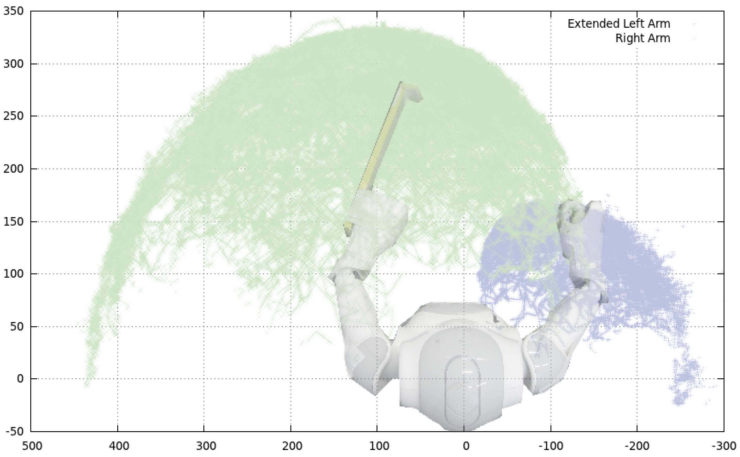


Figure 4.11: Reachable spaces for the arms of the Nao. The extension tool considerably modifies the action space of the left arm.

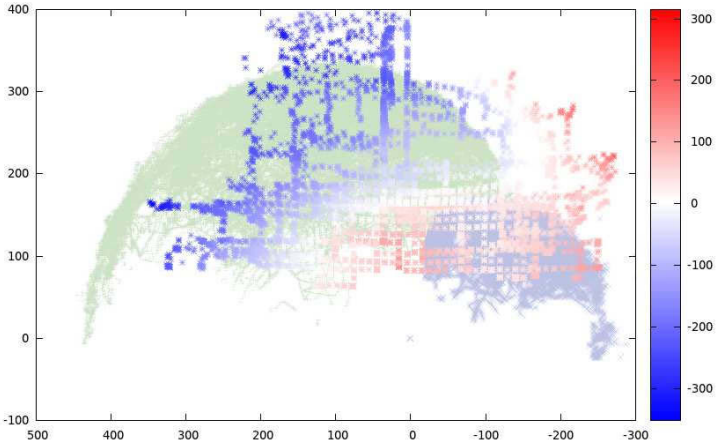


Figure 4.12: Prediction errors of the simulations of the left extended arm and the right arm movements.

4.3.3 Characterising self-produced movements

As discussed in the previous sections, forward models can be seen as self-monitoring mechanisms in humans. (Blakemore et al., 1998b) claimed that the sensory consequences of self-generated actions are perceived differently from an identical sensory input that is externally generated. This would explain the cancellation or the attenuation of tickle sensation when this is the consequence of self-generated motor commands. The capability to predict the sensory consequences of our own actions is fundamental in many basic motor tasks. Forward models, by functioning with self-generated motor commands are an important base for the feeling of agency (Gentsch and Schütz-Bosbach, 2011). Further evidence comes from studies where the motor or action based self model play a crucial role for recognition of self and others. In (Casile and Giese, 2006), it has been shown that we are better at recognising ourselves than others when watching movies of only point-light walkers. As we rarely see our selves performing this sort of action, inverse-forward models are thought to be involved in this type of recognition.

(Loula et al., 2005) studied the visual sensitivity in humans to biological movements. In particular, the authors tested the hypothesis that (1) if motor experience influences the visual analysis of action, then observers should be most sensitive to their own movements, and (2) if view-dependent visual experience determines visual sensitivity to human movement, then observers should be most sensitive to the movements of their friends. Results showed that sensitivity to one's own motion was highest and visual sensitivity to friends', but not strangers', actions was above chance.

(Rochat, 1998) showed that, in the course of the first weeks of life, infants develop an ability to detect intermodal invariants and regularities in their sensorimotor experience, which allow them to specify themselves as separate entities in the environment. Human biological motion has specific properties that makes it different from other types of motions (Blake and Shiffrar, 2007). For example, when people move their hands from one point to another, they often follow a straight line with a bell-shaped velocity profile, characterised by an initial acceleration phase followed by a longer deceleration phase (Abend et al., 1982). Another well-known property of human action is that movement velocity systematically varies with the curvature of the trajectory. In particular, the velocity of execution increases with the radius of the curvature of the trajectory (Lacquaniti et al., 1983).

We continued the study presented in (Schillaci et al., 2013c) by investigating how self-produced movements can be characterised in robots through the use of internal models[2]. In particular, we investigated whether the internal models representation can code invariant characteristics of the robot's movements. The rationale behind the investigation is that the internal simulations capabilities provided by such internal models could make robots sensitive to robotic motion, as well as humans are sensible to biological motion. We tried to replicate studies on biological motion, such as (Lacquaniti et al., 1983) and (Zwickel et al., 2012). For example, (Zwickel et al., 2012) showed that participants' performance in ocular tracking of point-light motions was better for biological than for non-biological (such as constant) motions. This would confirm that the human visual system is sensitive

[2]The experiment presented in this section has been co-authored by the same authors as in (Schillaci et al., 2013c).

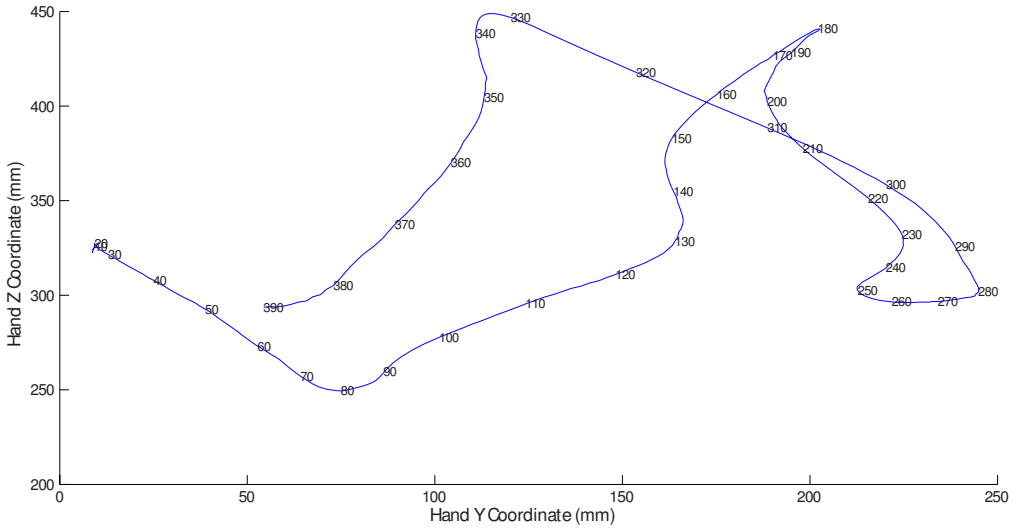


Figure 4.13: A sequence of typical babbling trajectories from the Aldebaran Nao robot, using the motion engine provided by the NaoTH framework. Here, only the projection of the trajectories into the Y-Z space is displayed. Numbers on the trajectory represent timestamps, expressed in frame numbers.

to biological motion (Blake and Shiffrar, 2007). As discussed in (Schillaci et al., 2013c), even when presented with point-light displays of human movement for which bodily form information is essentially absent, people can still rely on the available kinematic information to successfully track the motion (de'Sperati and Viviani, 1997), predict its outcome (Knoblich and Flach, 2001), or even recognise who produced it (Loula et al., 2005). Such effects generally lead to a so-called *self-advantage*: performance is better when perceiving one's own movements.

In a preliminary experiment, we programmed the humanoid robot Aldebaran Nao to gather sensorimotor experience through motor babbling. Motor babbling resulted in generating hand trajectories as the ones depicted in Figure 4.13. Following the hypothesis behind the studies presented in (Lacquaniti et al., 1983) and (Zwickel et al., 2012), we tried to make the robot's visual system sensible to robotic motion. Thus, an inverse and forward models pair has been trained using babbling sensorimotor data. In particular, as sensory states only the visual information represented by the $[x, y, z]$ coordinates of the robot's hand has been used. As motor commands, joints displacements have been considered.

Here, the robot observes trajectories with robotic and non-robotic (constant) velocity profiles. While observing the trajectories, internal simulations of the sensorimotor loop are run. In particular, the currently observed hand position and the previous one are fed into the inverse model, which generates a simulated motor command (or a joints displacement). Such an efferent copy of the motor command is sent to the forward model which

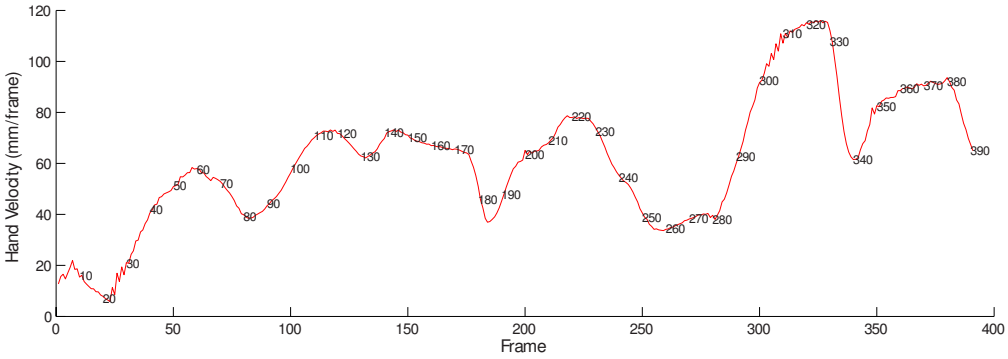


Figure 4.14: This graph illustrates how the robot's hand velocity varies along the trajectory showed in the previous Figure. The typical velocity profile of the robot is similar to the bell-shaped biological one. For example, timestamps 20 and 80 mark the endpoints of a rectilinear, to some extents, trajectory, as evident in the previous Figure.

outputs a predicted hand position. Instant by instant, a prediction error is calculated as the distance between the current hand position and the predicted one. The hypothesis behind the experiment is that the performance in predicting the hand trajectory is better when the robot observes a trajectory with robotic velocity profile than when it observes a trajectory with non-robotic velocity profile. However, in the preliminary experiment, the average prediction errors did not show any statistically significant difference, suggesting that the current configuration of the internal models is not properly coding the invariant characteristic of the robotic motion, in this case the velocity profile. Although the studies on biological motion have not been replicated successfully in this experiments, we found interesting properties of the used internal models configuration for a self-other distinction experiment, as illustrated in the following section.

4.3.4 Self-other distinction

Distinguishing between self and other is a cognitive ability that requires a basic understanding of our self and how we interact with the world (Schillaci et al., 2013c). To achieve this, we seem to rely on very finely tuned models of our motor capabilities. These models are involved in the control of our actions as well as in the prediction of the sensory consequences these have on our bodies and the environment (Casile and Giese, 2006). These predictions are what is thought to underlie our sense of ownership, and thereby provides us with a mean to recognise when actions are performed by others (Blakemore et al., 1998a).

In order to have robots interact naturally and intuitively with other agents, it is important that they first be able to recognise their own actions. In (Schillaci et al., 2013c), we presented an account on self-other distinction based on mental simulations of hand trajectories. Two experiments have been running. The aim of the first experiment was to test whether internal simulations can be used in distinguishing between hand trajectories produced by the Aldebaran Nao robot itself or by another type of robot, in this

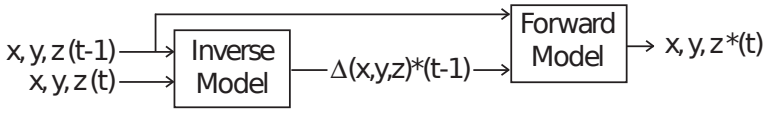


Figure 4.15: The inverse forward models pair used in the first experiment on self-other distinction. Only visual information has been used for coding sensory states (hand positions) and motor commands (hand displacements).

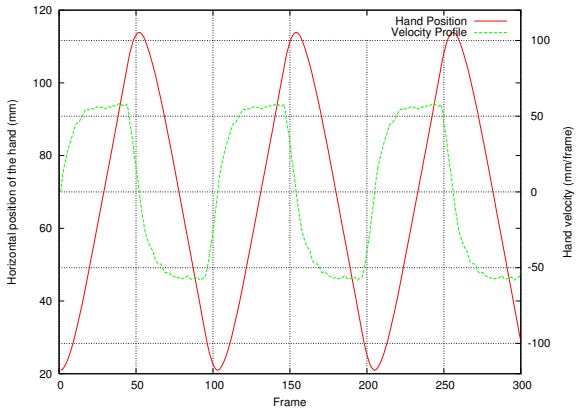


Figure 4.16: The red line depicts the trajectory of the Aldebaran Nao's hand during the performance test. The green line depicts the corresponding velocity.

case the robotic arm manipulator Puma. Firstly, as discussed before, the robot acquires sensorimotor experience through motor babbling. Thus, it codes the gathered data (only visual information has been used) into the internal models representation. Finally, predictive performances are compared from observing similar trajectories produced by the two different robots. In the second experiment, the internal models have been refined by adding proprioceptive information to the visual information and by comparing the robot movements with movements guided by a human experimenter.

Figure 4.15 shows the internal models configuration adopted in the first experiment. The babbling is performed in the joint space of the robot's arm, similarly to the experiment presented in the previous chapter but with larger movements (such as in the Purely Random (PR) movement strategy described in Appendix A). For each movement, the trajectory of the $3D$ position of the robot's hand has been stored. The babbling training session resulted in 41502 collected samples (each sample corresponds to a hand position). Each model is coded as a Multi-Layer Perceptron network. A sensory state is defined as the $[x, y, z]$ position of the hand. In this first experiment, a motor command is a displacement in the $[x, y, z]$ coordinates. Thus, the output of the inverse model is the necessary change in $3D$ coordinates to go from S_{t-1} to S_t. The simulated motor command $M*_{t-1}$ and the sensory situation S_{t-1} are used as input to the forward model, which predicts the resulting sensory situation.

The forward and inverse models have been coded as MLPs with 6 input neurons, 10 neurons in the hidden layer, and 3 output neurons. During training, the epsilon threshold

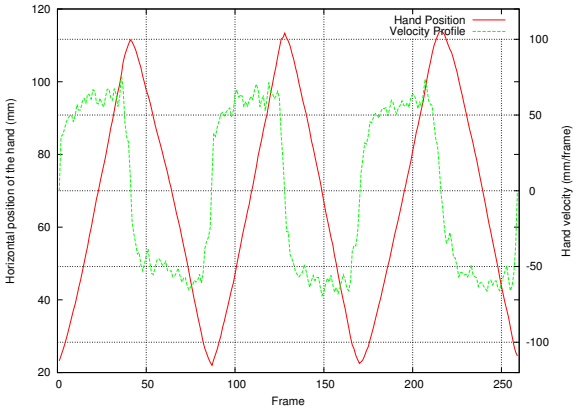


Figure 4.17: The red line depicts the trajectory of the Puma's hand during the performance test. The green line depicts the corresponding velocity.

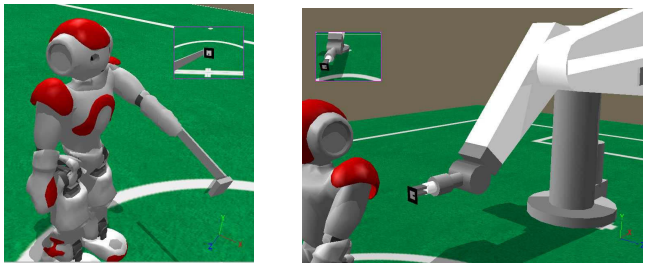


Figure 4.18: Experimental Setup: both the babbling training session for data collection and the testing session were run in the Webots robot simulator.

term criteria was reached after 18 iterations for the forward model and after 154 iterations for the inverse model.

Figures 4.16 and 4.17 show the execution of a number of controlled back and forth movements (along the $X-$axis) performed by the Aldebaran Nao robot and by the Puma robotic arm in the robot simulator Webots. These movements have distinctive trajectories (red lines) and velocity profiles (green lines) and we relied on internal simulations to distinguish between these two types of movements. Figure 4.18 shows two frames of the experimental setup.

Simulations of the sensorimotor loop were run by feeding the internal models with the sensory states taken from the two trajectories (Nao and Puma). Frame by frame, prediction errors were computed as the euclidean distance between the actual hand positions and the predicted ones. The prediction errors at each frame over 3 movement cycles were submitted to a two-tailed independent-samples t test, yielding a significant effect, $t(563) = 4.55$, $p < .001$. This result reflects that the mean prediction error for the Nao profile ($31.19mm$) was smaller than for the Puma ($35.74mm$) profile and is consistent with the self-advantage typically observed in prediction (Knoblich and Flach, 2001) and recognition (Loula et al., 2005) experiments involving humans.

In a second experiment[3], prediction performances have been tested on two types of observed movements: one performed by the Aldebaran Nao and one performed by the human experimenter, who manually guided the robot arm (see Figure 4.19). With manually guiding the robot arm (kinesthetic teaching), the human motor system is generating the trajectory. Thus, it is reasonable to think that the robot's original motion is different from the motion produced through kinesthetic teaching. The internal models representation has been refined compared to the one used in the Aldebaran Nao-Puma experiment. Here, sensory states are composed by 7 variables: $[x, y, z]$ position of the hand and the angles positions of the four arm joints (shoulder pitch, shoulder roll, elbow yaw, elbow roll). Motor commands are composed by 4 variables: velocities of the four arm joints (shoulder pitch, shoulder roll, elbow yaw, elbow roll). More information about the past has been included in the inverse and forward models. In particular, three configurations have been tested: (A) only one past sensory state (Figure 4.20), (B) two past sensory states (Figure 4.21) and (C) four past sensory states (Figure 4.22).

The babbling is performed in the joint space of the robot's arm. The training session resulted in ca. 43997 collected samples. Each model is coded as a Multi-Layer Perceptron network. In configuration (A), the inverse model consisted in an MLP with 14 nodes in the input layer (7 for the past sensory state and 7 for the current sensory state), 19 in the hidden layer and 4 output nodes; the forward model consisted in an MLP with 11 nodes in the input layer (7 for the previous sensory state and 4 for the motor command), 19 in the hidden layer and 7 output nodes. In configuration (B), the inverse model consisted in an MLP with 21 nodes in the input layer (14 for the past sensory state and 7 for the current sensory state), 26 in the hidden layer and 4 output nodes; the forward model consisted in an MLP with 18 nodes in the input layer (14 for the previous sensory state and 4 for the motor command), 26 in the hidden layer and 7 output nodes. In configuration (C), the inverse model consisted in an MLP with 35 nodes in the input layer (28 for the past sensory state and 7 for the current sensory state), 40 in the hidden layer and 4 output nodes; the forward model consisted in an MLP with 32 nodes in the input layer (28 for the previous sensory state and 4 for the motor command), 40 in the hidden layer and 7 output nodes.

The prediction performance of the inverse-forward models pair trained with robot babbling data has been compared with the one of the pair trained with human babbling data (that is, babbling movements produced by the experimenter with manually guiding the robot's arm). The human babbling resulted in 43376 collected samples, which have been used for training three different configurations of internal models, as for the robot data ((A) only one past sensory state (Figure 4.20), (B) two past sensory states (Figure 4.21) and (C) four past sensory states (Figure 4.22).

Simulations from the two competing pairs have been run while observing 6 trajectories, each of the duration of ca. 30 seconds: 3 performed by the robot itself and 3 performed by the human manually guiding the robot. Prediction errors at each frame for each trajectory were submitted to a two-tailed independent-samples t-test. Table 4.1 shows the significant effects and the mean prediction errors of the two internal models pairs (robot and human)

[3]Results still unpublished. Experiments co-authored by the same authors as in (Schillaci et al., 2013c).

Figure 4.19: Kinesthetic teaching

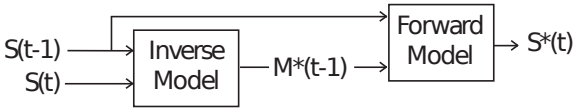


Figure 4.20: The internal models pair used in the second experiment on self-other distinction with one past sensory states. (Configuration A)

in configuration (A) while simulating the 6 trajectories. Table 4.2 shows the significant effects and the mean prediction errors of the two internal models pairs (robot and human) in configuration (B) while simulating the 6 trajectories. Table 4.3 shows the significant effects and the mean prediction errors of the two internal models pairs (robot and human) in configuration (C) while simulating the 6 trajectories.

Table 4.1: Mean prediction errors of the two internal models pairs (robot and human) in configuration (A) (only one past sensory state in the models) while simulating the 6 trajectories.

	Mean pred. err.(Robot)	Mean pred. err.(Human)	t-test sign.
robot trajectory 1	**0.000612323**	0.001230746	4.4365E-225
robot trajectory 2	**0.000871504**	0.001413089	1.1279E-228
robot trajectory 3	**0.000374848**	0.000964372	0
human trajectory 1	0.006669483	**0.000671424**	0
human trajectory 2	0.006376223	**0.000662538**	0
human trajectory 3	0.005777802	**0.000589336**	0

As shown in the tables, the robot internal models predicted the observed robot trajectories better than how the human internal models did, for all the three internal models configurations (only one past sensory state, two past sensory states and four past sensory states). Similarly, the human internal models predicted the observed human trajectories better than how the robot internal models did, for all the three internal models configurations. This result is consistent with the previous Aldebaran Nao-Puma comparison and

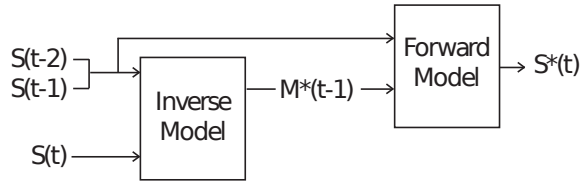


Figure 4.21: The internal models pair used in the second experiment on self-other distinction with two past sensory states. (Configuration B)

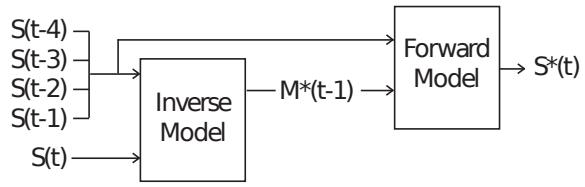


Figure 4.22: The internal models pair used in the second experiment on self-other distinction with four past sensory states. (Configuration C)

with the self-advantage typically observed in prediction (Knoblich and Flach, 2001) and recognition Loula et al. (2005) experiments involving humans.

4.3.5 Interacting with objects

Infants acquire governance and coordination of motor capabilities through interacting with the external world. In his theory of cognitive development, Piaget proposed that infants of 4 months of age start to become more object-oriented, moving beyond self-preoccupation. Actions involving interactions with objects bring interesting or pleasurable feelings to the baby.

Developmental mechanisms towards object-oriented actions have been implemented in robots by (Sheldon, 2012). Here, a robotic experiment is presented, where the humanoid robot Aldebaran Nao learns how to interact with an object through self-exploration and through observing a human demonstrator. Once having acquired this capability, the robot has been trained to recognise motor actions on objects. For simplicity, only three simple motor actions involving objects are addressed: reach an object, displace the object, withdraw the hand from the object. Grasping behaviours are not considered, here. Rather, only motor behaviours that can be characterised by the relationships between object and hand positions have been addressed. As in the previous experiments, internal models have been used for coding the gathered sensorimotor experience and internal simulation mechanisms have been used for implementing motor behaviour recognition capabilities.

The recent discovery of mirror neurons in the central nervous system supports the general idea of internal simulations. The mirror neuron system (MNS) is thought to be involved in internal simulations of the sensorimotor loop in learning and planning, as it has been found that neurons in this area show activation both when an individual performs a specific action and when the individual observes the same action performed by

Table 4.2: Mean prediction errors of the two internal models pairs (robot and human) in configuration (B) (two past sensory states in the models) while simulating the 6 trajectories.

	Mean pred. err.(Robot)	Mean pred. err.(Human)	t-test sign.
robot trajectory 1	**0.000525744**	0.001444192	8.6764E-227
robot trajectory 2	**0.000784814**	0.002035539	0
robot trajectory 3	**0.000514584**	0.000723721	2.7669E-111
human trajectory 1	0.005698069	**0.000453421**	0
human trajectory 2	0.00533802	**0.000453857**	1E-323
human trajectory 3	0.005798841	**0.000629191**	0

Table 4.3: Mean prediction errors of the two internal models pairs (robot and human) in configuration (C) (four past sensory states in the models) while simulating the 6 trajectories.

	Mean pred. err.(Robot)	Mean pred. err.(Human)	t-test sign.
robot trajectory 1	**0.000542368**	0.000962089	3.5845E-092
robot trajectory 2	**0.000850665**	0.001154319	9.2739E-053
robot trajectory 3	**0.000305658**	0.000577288	0
human trajectory 1	0.005015112	**0.000550195**	2.0422E-312
human trajectory 2	0.004704619	**0.000576965**	0
human trajectory 3	0.004141911	**0.000473002**	0

a demonstrator (for a review, see (Gallese, 2007)).

It seems that an observer understands a demonstrated behaviour comparing a simulated execution of it with a set of primitives stored in its memory. Recently, internal models have been used to try to explain and model the functioning of mirror neuron systems given their intrinsic capability of translating sensory data into motor data (Metta et al., 2006). If people simulate others' actions, then how accurately an observer can predict an observed action should depend on how closely the action maps onto the observer's own motor repertoire. (Eskenazi et al., 2009) also support the idea that perception and action matching allow us to exploit already existing predictive mechanisms in the motor system to make sense of others' actions.

As described at the beginning of this chapter, internal models have been used in cognitive robotics for the execution and recognition of actions (Dearden, 2008). Here, a similar experiment is presented, where the robot acquires similar sensorimotor skills by self-exploration and by observing a demonstrator. This is in line with the developmental process that leads towards social learning.

Self-exploration and learning from demonstrations

In (Schillaci et al., 2012b), an internal models based architecture was used in an experiment on motor behaviour recognition. The experimental set-up consisted in a humanoid robot

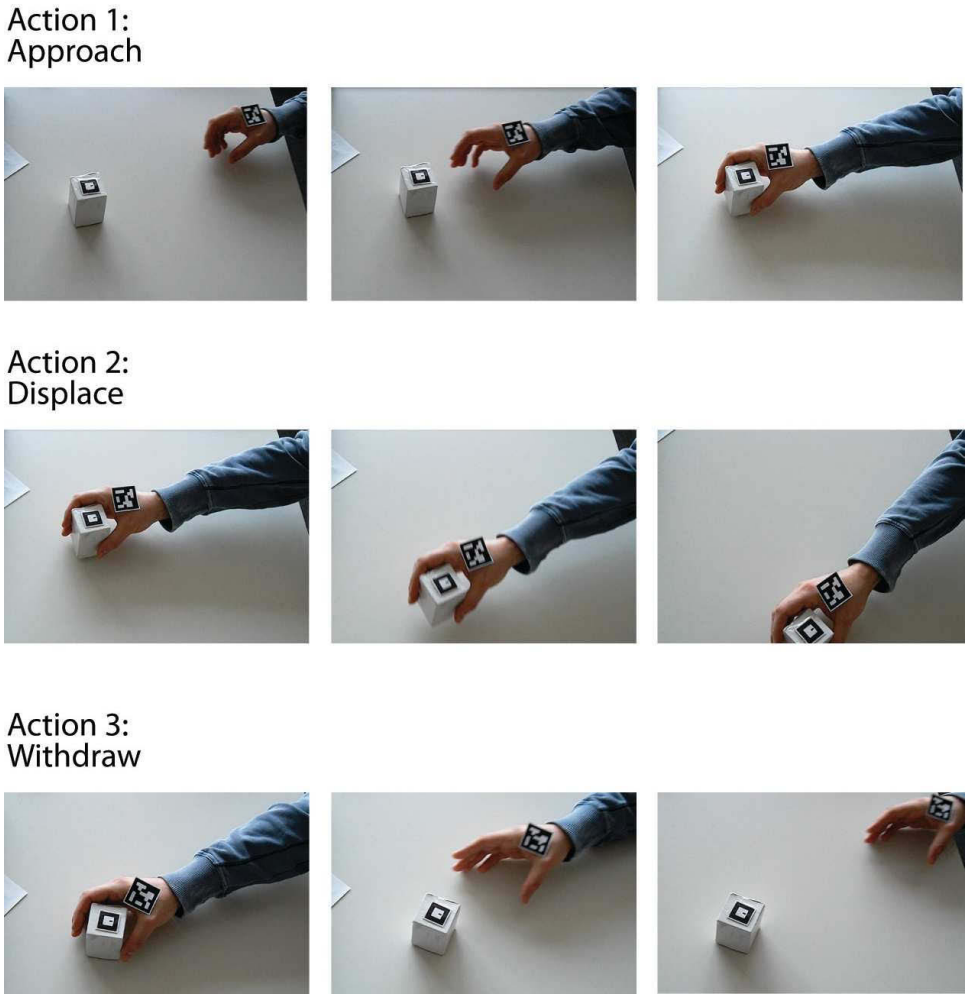


Figure 4.23: Typical human demonstrations of the three actions: approach the object, displace the object, withdraw the hand from the object.

observing a skilled demonstrator performing motor actions involving one or more objects, such as the ones shown in Figure 4.30. Training consisted in supervised learning by demonstration sessions, where multiple repetitions of the same action (such as approach or displace an object) were shown to the robot.

In the following sections, two studies on behaviour recognition are reported: an experiment where the robot has been programmed to learn by self-exploration three motor actions (approach, displace or withdraw the hand from an object) and an experiment (presented in (Schillaci et al., 2012b)) where actions have been learned from observing a human demonstrator.

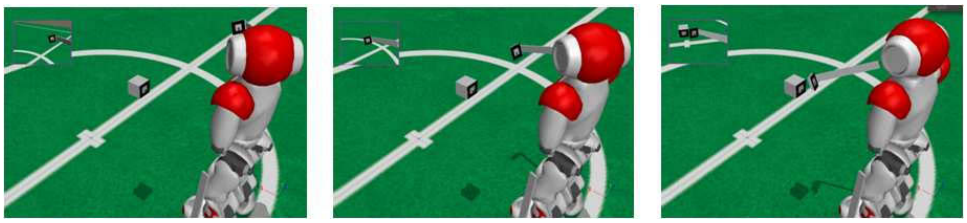


Figure 4.24: Typical babbling trajectory of the approach action. Motor babbling has been performed in Webots robotic simulator.

Recognising motor behaviours after learning from self-exploration

This section presents an experiment where the robot is programmed to learn by self-exploration the three motor actions: *approach*, *displace* and *withdraw* the hand from an object. Exploring the *approach* action consisted in repeating the following sequence of motor commands: (1) set the arm joints to a configuration which results in the end-effector being close to the object (starting configuration); (2) move the arm joints to a random configuration; (3) return to the starting configuration of the arm joints, using the original controller of the robot. While performing the third step, the robot is gazing at its end-effector and, at the same time, it is gathering sensory and motor data until it terminates the action. The resulting motor behaviour looks like, in fact, an approach action towards the object. Exploring the *withdraw* action consisted in performing the same sequence of steps as in the *approach* one, but recording the data during step (2). For learning the *displace* action, an object has been placed in the robot gripper and several iterations of step (2) have been generated, while the robot was gathering data from vision, proprioceptive sensors and motor commands.

Data gathered during the iteration of the three babbling behaviours have been used for training three inverse and forward models pairs (see Fig.4.25). Sensory states consisted of instances of 6 variables (d, $\delta(d)$, $j^{shoulderpitch}$, $j^{shoulderroll}$, $j^{elbowyaw}$, $j^{elbowroll}$), gathered from multi-modal sensory channels (vision and proprioception). In particular, d is the distance between hand and object (their 3D positions are estimated using fiducial markers); $\delta(d)$ is the derivative of the distance between hand and object; $j^{shoulderpitch}$ is the angle position of the shoulder pitch joint; $j^{shoulderroll}$ is the angle position of the shoulder roll joint; $j^{elbowyaw}$ is the angle position of the elbow yaw joint; $j^{elbowroll}$ is the angle position of the elbow roll joint.

Motor commands are instances of four variables: $\delta(j^{shoulderpitch})$, $\delta(j^{shoulderroll})$, $\delta(j^{elbowyaw})$, $\delta(j^{elbowroll})$. In particular, $\delta(j^{shoulderpitch})$ is the displacement of the shoulder pitch joint; $\delta(j^{shoulderroll})$ is the displacement of the shoulder roll joint; $\delta(j^{elbowyaw})$ is the displacement of the elbow yaw joint; $\delta(j^{elbowroll})$ is the displacement of the elbow roll joint. The displacement of each joint is measured as the angular distance covered by the joint in a constant time.

Table 4.4 illustrates the input and the output of each internal model in this behaviour recognition experiment.

Each of the three actions is characterised as a different tendency in the variations of

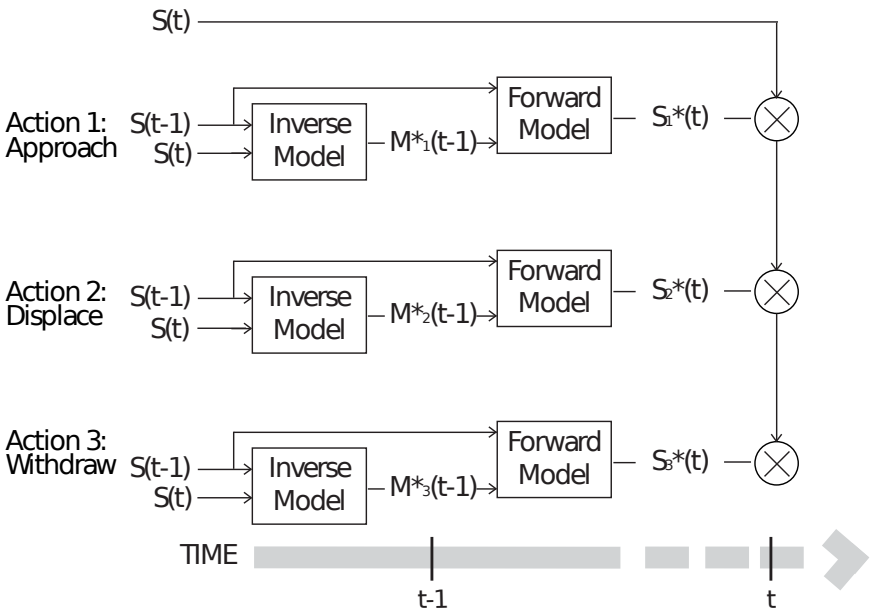


Figure 4.25: Competing inverse and forward models pairs for the behaviour recognition experiment.

such features. For example, the reach action is characterised by a decrease of the hand-object distance, thus a negative derivative of the distance. The withdraw action follows the opposite tendency: increase of the hand-object distance, thus positive derivative of the distance. The displace action is characterised by a constant value in this same variable. Including the data from the joints is fundamental for coding the robot's morphology and the characteristics of its motion profile.

During the motor babbling sessions, the robot was collecting, instant by instant, the following sensorimotor information: $[S_{t-1}; M_{t-1}; S_t]$. For each of the three actions (approach, displace and withdraw), the corresponding gathered sensorimotor data has been used in training two Multi-Layer Perceptrons:

- **Forward Model**: 10 nodes in the input layer (6 for S_{t-1} and 4 for M_{t-1}), 14 nodes in a single hidden layer, and 6 nodes in the output layer (S_t);
- **Inverse Model**: 12 nodes in the input layer (6 for S_{t-1} and 6 for S_t), 14 nodes in a single hidden layer, and 4 nodes in the output layer (M_{t-1});

A standard back-propagation algorithm has been used in training the MLPs[4].

During the motor babbling phase, 15227 samples have been collected for the approach action, 12951 samples for the displace one and 15679 for the withdraw one, with a frame

[4]Training parameters have been set to: Term criteria: MaxIteration=5000; Epsilon= 0.000001; Activation function = Symmetrical Sigmoid; dw-scale (the co-efficient to multiply the computed weight gradient by) = 0.05; moment-scale (the coefficient to multiply the difference between weights on the 2 previous iterations. This parameter provides some inertia to smooth the random fluctuations of the weights) = 0.05

Table 4.4: Input and output of the internal models.

Inverse Model			
Input	S_{t-1}	:	$d_{t-1}, \delta(d)_{t-1}, j_{t-1}^{shoulderpitch}, j_{t-1}^{shoulderroll}, j_{t-1}^{elbowyaw}, j_{t-1}^{elbowroll}$
	S_t	:	$d_t, \delta(d)_t, j_t^{shoulderpitch}, j_t^{shoulderroll}, j_t^{elbowyaw}, j_t^{elbowroll}$
Output	$M*_{t-1}$	:	$\delta(j^{shoulderpitch})_{t-1}, \delta(j^{shoulderroll})_{t-1}, \delta(j^{elbowyaw})_{t-1}, \delta(j^{elbowroll})_{t-1}$

Forward Model			
Input	S_{t-1}	:	$d_{t-1}, \delta(d)_{t-1}, j_{t-1}^{shoulderpitch}, j_{t-1}^{shoulderroll}, j_{t-1}^{elbowyaw}, j_{t-1}^{elbowroll}$
	M_{t-1}	:	$\delta(j^{shoulderpitch})_{t-1}, \delta(j^{shoulderroll})_{t-1}, \delta(j^{elbowyaw})_{t-1}, \delta(j^{elbowroll})_{t-1}$
Output	$S*_t$	:	$d_t, \delta(d)_t, j_t^{shoulderpitch}, j_t^{shoulderroll}, j_t^{elbowyaw}, j_t^{elbowroll}$

rate of ca. 25 fps. For the approach action, training the forward model took 68 iterations, reaching a squared error $E = 59.53$; training the inverse model took 69 iterations, reaching $E = 762.34$. For the displace action, training the forward model took 626 iterations with $E = 849.68$, while it took 2517 iterations for the inverse model, reaching $E = 2369.88$. For the withdraw action, training the forward model took 27 iterations, with $E = 311.82$, while for the inverse model it took 1106 iterations with $E = 241.63$.

Classification performances using internal simulation have been tested on sequences of random generated trajectories of the three actions (approach, displace, withdraw). The trajectories have been executed by the robot itself. Frame by frame, the robot estimates hand and object positions and records the proprioceptive information about its joint positions. With this data, it computes the sensory states S_{t-1} and S_t. Internal simulations of the sensorimotor loop are performed for each action. The simulated outcome S_t^* for each controller-predictor pair is compared with the actual sensory situation S_t. The action corresponding to the pair with the least error is chosen as the most probable observed one.

Internal simulations are performed as follows. For each controller-predictor pair, compute:

- inverse prediction: predict the motor command M_{t-1} from the sensory situations S_{t-1} and S_t;
- forward prediction: predict the sensory situation S_t^* generated by the predicted motor command (the outcome of the inverse model) applied to the previous sensory situation S_{t-1};
- error of the simulation: calculate the Mahalanobis distance between the actual sensory situation S_t and the predicted one S_t^*.

The pair with the lowest error is selected as the most probable observed action. Performing a recognition session of almost 2.5 minutes (153.96 seconds), where the robot is alternating the execution of the three actions, the system had a correct recognition rate of 78.40% (that is, 78.40% of the times the action corresponding to the controller-predictor pair with the lowest error was the one being observed). Test trajectories have been generated in two steps: firstly, a sequence of approach/withdraw actions have been generated

using the babbling algorithm adopted for exploring the approach/withdraw motor actions. Webots simulator have been used for generating the sensorimotor data. Trajectories look similar to the sequence illustrated in Figure 4.24, in going back and forth from random positions to the object one. Secondly, test trajectories of the displace actions have been generated using the mechanism adopted for babbling the same displace action. Choosing an action randomly would have resulted in a correct recognition rate of 33.33%.

Although a standard machine learning algorithm has been adopted for training the internal models, the high correct classification rate proves the power of the internal simulation paradigm. It has to be noted that the knowledge bases contained relatively clean information in the way that, due to the supervised learning, actions have been manually temporally segmented.

Table 4.5 shows the confusion matrix of the classifier, comparing the prediction outcomes with the actual outcomes. The confusion matrix indicates how much (in percentiles of the actual outcome) every demonstrated action has been recognised as approach, displace or withdraw action.

Table 4.5: Confusion Matrix of the behaviours classifier trained with self-exploration data

		Predicted Outcome		
		approach	displace	withdraw
Actual Outcome	approach	81.17%	0.00%	18.83%
	displace	8.02%	81.11%	10.87%
	withdraw	27.26%	0.00%	72.74%

Recognising motor behaviours after learning by demonstration

In (Schillaci et al., 2012b), a similar internal models based architecture has been used as a computational model for a behaviour recognition experiment. The experimental set-up consisted in a humanoid robot observing a skilled demonstrator performing motor actions similar to those described in the previous section: approach an object and displace an object. However, in the experiment presented here, the robot has to classify not only the observed action, but also the target of that action, since multiple objects are present in the scene.

Training consisted in supervised learning by demonstration phases, where multiple repetitions of the same action (approach and displace the object) were shown to the robot. For each action, a pair of inverse-forward models has been trained with information extracted from the movements of a demonstrator (motor commands) and from the relationship between the positions of its hand and the one of the object (sensory states).

As in the experiment presented in the previous section, (Schillaci et al., 2012b) showed how sensorimotor loops can be simulated during the observation of new actions demonstrations, and how internal simulations can produce prediction errors useful for classifying the observed behaviour.

Internal models have been trained with a different configuration of sensory states and motor commands than the one described in the previous section. In particular, sensory states included the following characteristics computed from the visual input: d, δ, θ and ϕ. d represents the distance between hand and object (their 3D positions are estimated using fiducial markers); δ represents the derivative of the distance between hand and object; θ and ϕ represent the orientation of the object with respect to the hand. This characteristic is coded as two angles, the latitude (θ) and the longitude (ϕ) of the object position in a frame of reference centered on the hand position.

In this experiment, the sensory situation is coded as an instance of the previous characteristics: d, δ, θ and ϕ. For each time step, the characteristics encoding such a sensory situation are calculated from the positions of the hand and the object. The motor command is coded as the three components of the velocity vector describing the movement of the hand: v^x, v^y and v^z. Table 4.6 illustrates the input and the output of each internal model in the behaviour recognition experiment.

Table 4.6: Input and output of the internal models.

Inverse Model			
Input	S_{t-1}	:	$d_{t-1}, \delta_{t-1}, \theta_{t-1}, \phi_{t-1}$
	S_t	:	$d_t, \delta_t, \theta_t, \phi_t$
Output	$M*_{t-1}$	:	$v^x_{t-1}, v^y_{t-1}, v^z_{t-1}$

Forward Model			
Input	S_{t-1}	:	$d_{t-1}, \delta_{t-1}, \theta_{t-1}, \phi_{t-1}$
	M_{t-1}	:	$v^x_{t-1}, v^y_{t-1}, v^z_{t-1}$
Output	$S*_t$	:	$d_t, \delta_t, \theta_t, \phi_t$

In (Schillaci et al., 2012b), two mechanisms have been adopted for coding the internal models and for running internal simulations: (A) a knowledge base, where sensorimotor experience has been stored and a k-Nearest Neighbours based algorithm has been used as inference tool for the simulations, and (B), as in the previous experiments, a Multi-Layer Perceptron.

Supervised learning sessions were performed off-line by using recorded videos. The robot observed demonstrations of each action, manually segmented by the user. In configuration (A), for each video, data represented by the characteristics specified before were collected into a knowledge base. Each component of the knowledge base, collected at time t, contains the following information: $[S_{t-1}; M_{t-1}; S_t]$, which means that at each time step the previous sensory situation S_{t-1}, the current sensory situation S_t and the motor command M_{t-1} that caused S_{t-1} to become S_t have been saved as an element of the knowledge base.

In (Schillaci et al., 2012a), a k-Nearest Neighbours based algorithm was used as inference tool for the inverse and forward predictions. For inverse model predictions, the motor command $M*_{t-1}$ which changes the sensory situation from S_{t-1} to S_t is calculated as follows: Given the hand and object positions at time $t-2$ and $t-1$, the features

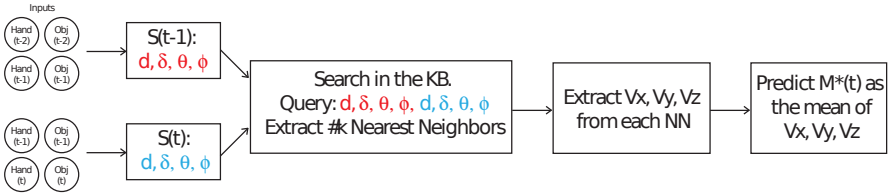


Figure 4.26: An illustration of the inverse model prediction with k-NN.

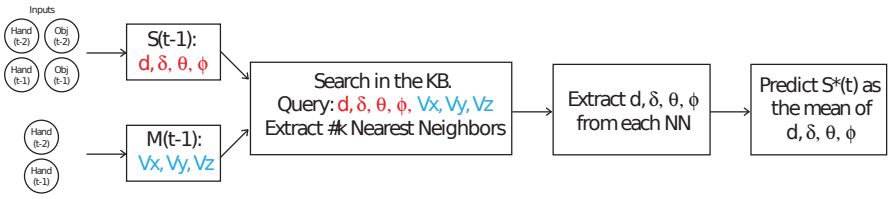


Figure 4.27: An illustration of the forward model prediction with k-NN.

which compose the sensory situation S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, are calculated[5]. In a similar way, given the hand and object positions at time $t-1$ and t, the features which compose the sensory situation S_t, i.e. d_t, δ_t, θ_t and ϕ_t, are calculated. A k-NN search in the knowledge base is then performed, where the query is composed by S_{t-1} and S_t. Finally, the M_{t-1} components, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}, are extracted from the k found vectors and their mean is the output of the inverse model prediction. Figure 4.26 illustrates the algorithm for performing inverse predictions using k-NN.

Similarly, the forward model predictions are calculated as follows: given the hand and object positions at time $t-2$ and $t-1$, the features which compose the sensory situation S_{t-1}, i.e. d_{t-1}, δ_{t-1}, θ_{t-1} and ϕ_{t-1}, are calculated. Then, given the hand positions at time $t-2$ and $t-1$, the motor command M_{t-1} is calculated as the derivative of the displacements in each direction, i.e. v^x_{t-1}, v^y_{t-1} and v^z_{t-1}. A k-NN search is performed, but now the query is composed by S_{t-1} and M_{t-1}. The final step consists in extracting from the k found vectors the S_t components, i.e. d_t, δ_t, θ_t and ϕ_t, and returning their mean as the forward model prediction. Figure 4.27 illustrates the algorithm for performing forward predictions using k-NN.

Classification performances have been tested in an action recognition experiment, where the robot is facing towards an action demonstration and is expected to recognise the observed action in real time. Frame by frame, it estimates hand and object positions and it computes both sensory states S_{t-1} and S_t. Internal simulations of the sensorimotor loop are performed for each action, that is for each controller-predictor pair. First, S_{t-1} and S_t are fed into the inverse model which predicts the motor command M^*_{t-1}; then, S_{t-1} and M^*_{t-1} are sent to the corresponding forward model to generate the simulated outcome S^*_t. Each of these predictions is then compared with the actual sensory situation S_t. The action corresponding to the pair with the least error is chosen as the most

[5]Hand and object positions at time $t-2$ are needed for estimating the sensory situation S_{t-1} (for example the variation of the hand-object distance between the current instant and the previous one).

probably observed one. Mahalanobis distance has been used for calculating the prediction error.

As stated before, two configurations have been used for coding the internal models: a knowledge base for k-NN search and an MLP. As in the experiments presented in the previous sections, in the second configuration, each inverse and forward model has been coded as a multi-layer perceptron which has been trained with a back-propagation algorithm using the data collected during the supervised learning sessions. The input and output nodes of the internal models are the same as in the k-NN case.

In the previous section, we performed internal simulations in order to estimate which inverse-forward model pair was more closely coding for the observed demonstration (where the robot itself was executing the action). This section shows how this system can be used to recognise not only the action performed on an object, but also the target object of the action, when two objects are present in the scene. As described before, the sensory states S_{t-1} and S_t of the internal models correspond to certain relationships between the position of the hand of the demonstrator and the one of the target object. During the demonstration, S_{t-1} and S_t are extracted from such positions and sent into the inverse and forward model to generate motor and state predictions. Each state prediction S_t^* (that is, the outcome of each controller-predictor pair) is then compared to the actual one, S_t. The observed behaviour is then classified as the one corresponding to the pair which results in the lowest prediction error.

In presence of multiple objects in the scene, simulations can be performed for each object-hand relationship. The same internal models can be fed with the states computed using the relationship between the position of the hand and each one of the objects, for example S^1_{t-1} (i.e. d^1_{t-1}, δ^1_{t-1}, θ^1_{t-1} and ϕ^1_{t-1}) and S^1_t (i.e. d^1_t, δ^1_t, θ^1_t and ϕ^1_t) as the states computed with the position of object 1, S^2_{t-1} and S^2_t as the states computed with the position of object 2, etc. Thus, each inverse-forward model pair can be fed with each of these couples of states and prediction errors can be computed with their corresponding desired states. In this way, the target object of the ongoing action can be inferred as the one which corresponds to the best inverse-forward model pair fed with the states computed with its position.

In (Schillaci et al., 2012b), two objects were present in the scene, namely object 1 and object 2, and two inverse-forward model pairs have been tested (the first coding for the *approach* action and the second coding for the *displace* action). The system computes the states S^1_{t-1} and S^1_t (using the position of the hand and the one of object 1) and the states S^2_{t-1} and S^2_t (using the position of the hand and the one of object 2). S^1_{t-1} and S^1_t are sent to the pair *approach*, a prediction S^{1*}_t is calculated and compared with the state S^1_t, resulting in the prediction error $ERR^1_{approach}$. In the same way, S^2_{t-1} and S^2_t are sent to the pair *approach*, resulting in the prediction error $ERR^2_{approach}$. The same process is done with the pair *displace*, resulting in two more prediction errors, so that in total we have: $ERR^1_{approach}$, $ERR^2_{approach}$, $ERR^1_{displace}$ and $ERR^2_{displace}$. The smallest error corresponds to the best pair which is fed with the data of the most probable target of the action.

In this experimental setup, two inverse-forward models pairs have been trained with data collected from the observation of two actions directed to an object: *approach* and

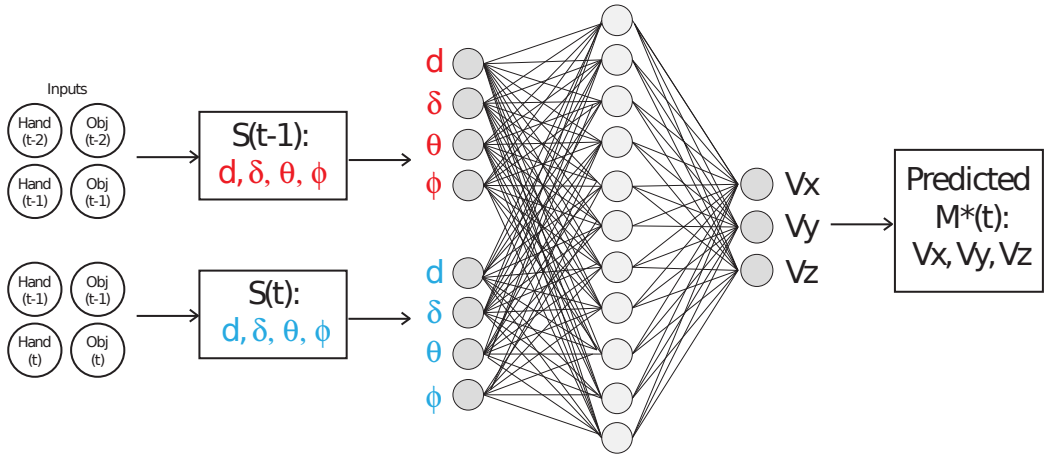


Figure 4.28: An illustration of the inverse model prediction with MLP.

displace. In particular, 1004 training samples have been gathered from 83 demonstrations of the approach action; 3245 training samples from 108 demonstrations of the displace action. Each sample contains $[S_{t-1}; M_{t-1}; St]$, where the state S is a 4-dimensional vector containing the same features described in the previous configuration using k-NN as inference tool: d, δ, θ and ϕ. As before, the motor command is a 3-dimensional vector representing v^x, v^y and v^z, the components of the hand velocity.

Thus, the prediction performances of two machine learning tools have been compared: k-Nearest Neighbours (three configurations: $k = 5$, $k = 11$, $k = 55$) and Multi-Layer Perceptrons. In the latter case, forward models have been coded as MLPs with 7 input neurons, 12 neurons in the hidden layer, and 4 output neurons[6]. Inverse models have been coded as MLPs with 8 input neurons, 16 neurons in one hidden layer and 3 output neurons (4 input neurons to code for S_{t-1} plus 4 input neurons for S_t and 3 output neurons for M_{t-1}). In training the internal models for the *approach* behaviour, the epsilon threshold term criteria has been reached after 437 iterations for the forward model and after 2136 iterations for the inverse one. In training the *displace* action, the epsilon threshold term criteria has been reached after 333 iterations for the forward model and after 1891 iterations for the inverse one. Figure 4.28 and 4.29 illustrate the algorithms for inverse and forward predictions using MLPs.

The following tables show the confusion matrices using four inference tools: MLP (Multi-Layer Perceptron), 5-NN (k-Nearest Neighbours with $k = 5$), 11-NN and 55-NN. The confusion matrix indicates how much (in percentiles of the actual outcome) every demonstrated action has been recognised as approach or displace with target object 1 or 2. The correct classification rates were: 89.45% for the MLP, 61.81% for the 5-NN, 63.82% for the 11-NN and 64.82% for the 55-NN, claiming MLP as the best performing

[6]Term criteria: MaxIteration=500000; Epsilon= 0.000001; Activation function = Symmetrical Sigmoid; Training algorithm = BackPropagation; dw-scale (the coefficient to multiply the computed weight gradient by) = 0.05; moment-scale (the coefficient to multiply the difference between weights on the 2 previous iterations. This parameter provides some inertia to smooth the random fluctuations of the weights) = 0.05

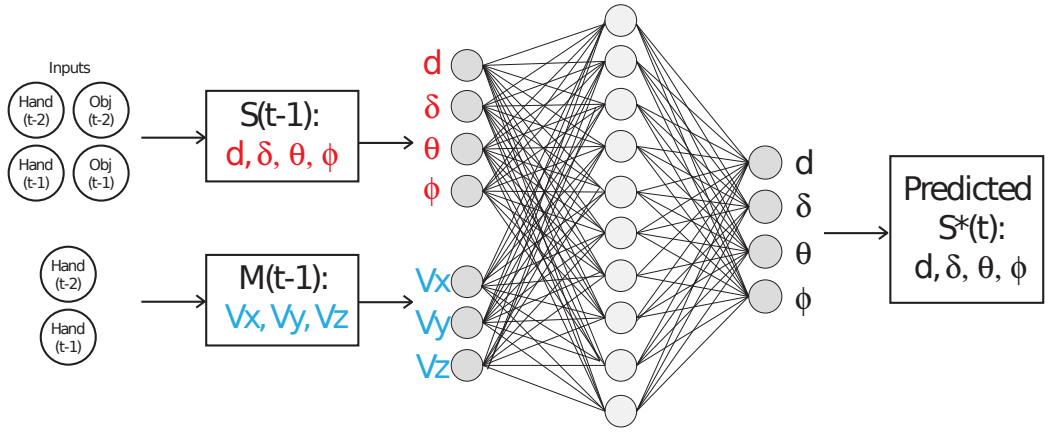


Figure 4.29: An illustration of the forward model prediction with MLP.

tool. Figure 4.30 shows a demonstration of the behaviour and target recognition using MLP.

Table 4.7: Confusion Matrix of the behaviours classifier trained with human demonstrations. Inference tool: MLP.

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	100.00%	0.00%	24.45%	7.25%
	displace-obj1	0.00%	95.52%	0.00%	0.00%
	approach-obj2	0.00%	4.48%	73.33%	0.00%
	displace-obj2	0.00%	0.00%	2.22%	92.75%

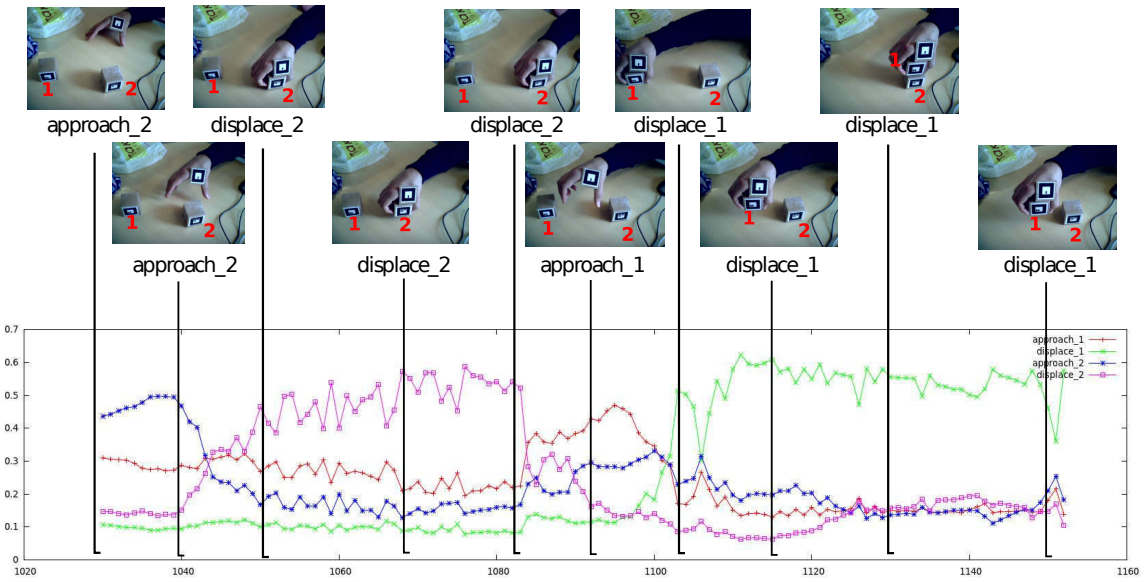


Figure 4.30: Demonstration of the behaviour and target recognition using MLP. The bottom graph shows the probabilities of each action to each object.

Table 4.8: Confusion Matrix of the behaviours classifier trained with human demonstrations. Inference tool: k-NN ($k = 5$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	57.78%	39.13%
	displace-obj1	0.00%	97.01%	4.44%	0.00%
	approach-obj2	29.41%	2.99%	8.89%	0.00%
	displace-obj2	0.00%	0.00%	28.89%	60.87%

Table 4.9: Confusion Matrix of the behaviours classifier trained with human demonstrations. Inference tool: k-NN ($k = 11$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	55.56%	33.33%
	displace-obj1	0.00%	95.52%	2.22%	0.00%
	approach-obj2	29.41%	4.48%	11.11%	0.00%
	displace-obj2	0.00%	0.00%	31.11%	66.67%

Table 4.10: Confusion Matrix of the behaviours classifier trained with human demonstrations. Inference tool: k-NN ($k = 55$).

		Actual Outcome			
		approach-obj1	displace-obj1	approach-obj2	displace-obj2
Prediction Outcome	approach-obj1	70.59%	0.00%	51.11%	31.88%
	displace-obj1	0.00%	97.01%	2.22%	0.00%
	approach-obj2	29.41%	2.99%	11.11%	0.00%
	displace-obj2	0.00%	0.00%	35.56%	68.12%

Chapter 5

Conclusions

This thesis began with three research questions in the context of autonomous motor and mental development in artificial systems:

1. What are the basic behavioural components an artificial agent should be provided with for being able to develop motor and cognitive capabilities?

2. How can an artificial agent represent and store the experience generated through such basic behaviours?

3. How can the acquired experience be reused and what computational processes are needed for generating basic cognitive skills out of it?

In approaching these questions, three assumptions have been made. Firstly, it is assumed that cognition is rooted in the bodily experience with the world. Thus, when implementing cognitive capabilities in artificial systems, it is necessary to study the agent *with a body*, which develop motor and cognitive capabilities through interacting with the environment. This thesis investigates sensorimotor interactions as a mean for the acquisition of experience in robots, where such interactions are resulting from the agents' bodily characteristics. Secondly, in line with theories on grounded cognition, it is assumed that mental simulation of sensorimotor experience could be a form of computation that could account for some of the cognitive processes that occur *offline*, that is in absence of external stimuli. The third assumption is that studying human development could give insights in finding those basic behavioural components that may allow for the autonomous mental and motor development in artificial agents.

In finding the answer to the first research question, studies on human development and developmental robotics have been reviewed. It has been argued that embodiment is a crucial factor to take into account when implementing cognitive skills in artificial agents. It has been pointed out that defining models of robots' embodiment and their surrounding world *a priori* should be avoided, since the risk is to stumble across problems such as robot behaviours lacking adaptability and capability to react to unexpected circumstances. Therefore, a developmental approach is needed, where artificial agents

are provided with mechanisms based on interactions with the physical and social environment, through which they can develop increasingly more complex motor and cognitive capabilities and become more autonomous, adaptable and social (Lungarella et al., 2003). In chapter 3 of this thesis, a number of basic behavioural components an artificial agent should be provided with for being able to develop motor and cognitive capabilities have been identified. The skills have been elicited along two developmental timelines, *sensorimotor* and *social*. At the origin of the two timelines, a basic behavioural component has been identified, namely body babbling, which allows an artificial agent to autonomously acquire sensorimotor experience by self-exploration. Through exploratory behaviours, infants acquire governance and coordination of their motor capabilities, which are needed to understand the objects they are surrounded by, to recognise them and to play with them. This thesis is inspired by such a developmental progression and it presents experiments on exploration behaviours for mapping different sensorimotor modalities in artificial agents, on the exploration of arm movements capabilities, on exploration under particular situations where the morphology of the robot is altered (such as with the usage of a tool) and on the exploration of sensorimotor behaviours involving interactions with objects.

The development of social skills has been also addressed. In particular, the last sections of chapter 3 are concerned with the topic of joint attention. In developmental psychology, several studies demonstrated that the development of skills for sharing attention between individuals lays the foundation of imitation learning and social cognition. However, how to implement such capabilities in robots is still an open challenge. Two prerequisites of joint attention have been investigated here: attention manipulation skills through pointing gestures and mechanisms for detecting visually salient events such as movements, faces or objects. In particular, it has been studied how proto-imperative pointing gestures can emerge from failed grasping actions and how attention manipulation and visual saliency detection mechanisms integrated with a short-memory system based on a robot ego-sphere can affect human-robot interaction.

Regarding the second research question, the debate on how the brain stores experience is still open. Theorists on embodied and grounded cognition converge on the idea that experience is stored in the form of multi-modal sensorimotor representations. Such theories are supported by several studies in behavioural and cognitive sciences. Thus, in Chapter 3, mechanisms for the autonomous acquisition of *sensorimotor* experience in artificial agents have been addressed. In fact, sensorimotor control is a remarkable capability of natural systems in their interaction with the environment. In the human developmental timeline, sensorimotor skills are acquired and rehearsed through a long process of interaction with the environment and with other subjects. In this thesis, exploration behaviours based on sensorimotor coordination have been studied for allowing a robot to gather multi-modal experience about its corporeality. Therefore, the established framework of the internal models has been adopted, since it allows for coding *sensorimotor* experience in artificial agents. Internal models, especially inverse and forward models, have not been fully studied and their usefulness not properly exploited in robotics. Forward and inverse models become central players in cognition, as they naturally fuse together different sensory modalities as well as motor information providing agents with multi-modal representations. Subjective experience can be coded into the internal models representation. This

thesis contributes with a deeper investigation of internal models by testing them in different experimental setups. In particular, inverse and forward models have been trained with different configurations of low-level sensory and motor data generated by the robot through exploration behaviours, demonstrated by a human, or acquired through kinesthetic teaching. As discussed in Chapter 2 of this thesis, this approach is opposed to the classical GOFAI's one, where knowledge is composed by a-modal representations and intelligence is seen as a manipulation of these symbolic representations of the world.

In dealing with the third research question, studies on mental imagery and predictive mechanisms in humans have been studied. In fact, it is plausible that simulations of sensorimotor cycles are behind some of the basic cognitive mechanisms that occur off-line, that is, in the absence of actual perceptual experience. Many studies suggested that sensorimotor prediction processes exist in motor planning and execution, in motor behaviour recognition, in control of speech, in mental rotation tasks, in self-monitoring mechanisms and perhaps in self-other distinction capabilities. This thesis investigated how basic cognitive skills can be implemented in a humanoid robot by allowing it to reuse the knowledge gathered in the past interactions with the world. In particular, mechanisms for generating recreations of perceptual and motor experience have been implemented, namely internal simulations. These processes have been used for implementing basic cognitive skills in a humanoid robot, such as action selection, tool-use, behaviour recognition and self-other distinction.

Differently from most of the existing robotics architectures implementing the internal models paradigm, this thesis tried to make the least assumptions possible in the definition of the functions modelling the controller-predictor pairs. For example, relevant studies often leave aside the learning of low level controllers, assuming that the inverse models are already present in the motor repertoire of the artificial agent.

On the one hand, this thesis' contribution aims at filling this gap, focusing on learning both the inverse and the forward models at the lowest possible level. In fact, efforts have been spent on finding minimal configurations of the sensorimotor information necessary for the emergence of the cognitive capabilities shown in the experiments. In the study on the characterisation of robotic motion, different configurations of sensorimotor variables have been tested. The aim was to find invariant characteristics of the robotic motion profile using basic information, such as hand position and arm joints displacements. In addition, in the self-other distinction experiment, the velocity profile of a robot controller has been coded into an internal models representation, by using only visual information, such as hand velocity. In the behaviour recognition experiment, not only the robot has been programmed to learn motor actions both by self-exploration and from observing a human demonstrator, but also several configurations of sensorimotor data have been tested in training the internal models, including visual information, such as hand/object positions and their relationships, and proprioceptive data, such as arm joints displacements.

On the other hand, the importance of self-exploration in the definition of the motor based self-model and of subjective experience has been pointed out. The use of basic and low level sensorimotor data should allow roboticists the discovery and therefore exploitation of properties that characterise the particular platform performing the learning. In the experiments presented here, the robot acquires sensorimotor skills by exploring its

action space, by being manually guided or by observing a human demonstrator. This is in line with the developmental process that leads towards social learning. Moreover, internal models have been shown to be capable of modelling experience gathered from the three learning mechanisms.

To conclude, this thesis aims at advancing the state-of-the-art in the cognitive robotics community with new demonstrations of the potentialities of sensorimotor learning, of the internal models framework, of internal simulation processes and of attentive mechanisms in the development of cognition. Applying the methods described in this work can improve the quality of robot capabilities for learning, adapting, reacting to unexpected circumstances, exhibiting a proper level of intelligence and autonomously and safely operating in unconstrained and uncertain environments. Such robots could be adopted in different application domains: from service domains, such as surveillance, inspecting and renovating in hazardous environments, agriculture, firefighting, medicine, floor cleaning - to personal and social domains, such as entertainment, elderly and nursing care, autism therapy, rehabilitation, and so on.

5.1 Outlook

This work can be extended towards many directions. Of particular importance is continuing the investigation on self-recognition and agency. A robot aware of itself and of its own actions can better interact with humans in a real world. Multi-modal sensorimotor representations should allow a robot the internalisation of a sensorimotor model of itself, its own actions and a differentiation with those of others. A research direction could be the characterisation of robotics motion: finding invariant components in the sensorimotor information that the robot experiences during the interaction with the world, can allow it to model itself and its own movements as a particular entity in the world, different than others.

Generation of body maps and their plasticity is a very important research topic that can be further explored. For example, investigating the capability to use tools could provide promising results in robot manipulators. On the one hand, it is grounded on complex sensorimotor transformations, thus it would involve research on more complex representations. On the other hand, research on tool-use could give insights on robot capabilities to acquire such rules and to re-adapt to morphological changes (Bongard et al., 2006).

Another research direction refers to active learning and curiosity-driven exploratory mechanisms. In fact, although human foetuses apparently explore motor capabilities in a random fashion, it has been shown by (von Hofsten, 1982) that babbling (in the case of reaching) in few days old newborns seems to be goal-directed. Von Hofsten showed that infants produce more arm movements toward an object, when it is in the visual field, than movements away from it. (Rolf et al., 2011) pointed out that this indicates a strong role of "learning by doing" instead of random exploration and that infants learn to reach by trying to reach. In artificial systems, exploration behaviours such as motor babbling have been traditionally implemented as random motor commands. Recently, a new trend is

pushing towards implementing curiosity-driven and active learning mechanisms. In fact, as noted by (Rolf et al., 2011), random exploration becomes very inefficient with increasing dimensionality of the sensorimotor space. The exploration can be significantly improved by active learning schemes. In (Rolf and Steil, 2012), the authors investigate goal-directed behaviours for exploring high dimensional sensorimotor spaces and for learning of inverse kinematic models. Their results have been shown to outperforms previous exploration approaches by several orders of magnitude. Other researchers confirmed such results, such as (Saegusa et al., 2009).

It is of great importance deepening the investigation on the development of joint attention and social skills in robots. Mechanisms similar to those presented in this work for acquiring imperative pointing gestures capabilities could be adopted, for instance, for the development of eye-gaze following. Of particular interest is also investigating whether eye-gaze following and pointing gestures share the same meaning and, if so, how such a shared meaning emerges. The saliency detection mechanisms could be extended by including additional saliency maps, such as auditory maps. Behaviours resulting from visual and auditory saliency detection mechanisms could be integrated with a more complex exploratory system for the generation of richer sensorimotor information.

Finally, extending the internal models framework could provide robots with richer motor and cognitive capabilities. For instance, Self-Organising Maps could be adopted in coding multi-modal channels, such as in the Epigenetic Robotics Architecture proposed by (Morse et al., 2010). On-line and adaptive training algorithm could be used for training the inverse and forward models. A richer representation of the models could be studied, e.g. for allowing hierarchical representations and for including motion primitives as motor commands information.

Appendix A

Random Movement Strategies

This appendix presents the results of the study conducted in (Schillaci and Hafner, 2011a) and (Schillaci and Hafner, 2011b) about random movement strategies for self-exploration in a humanoid robot. Motor babbling is adopted as the learning strategies of a humanoid robot that maps its random arm movements with its head movements, determined by the perception of its own body. The humanoid robot Aldebaran Nao is equipped with an elementary attentive system for perceiving its own body and for moving its head to focus on it (see Figure 3.2 in Chapter 3). In the cited articles, three random movement strategies have been implemented on a humanoid robot and their performance about the learning speed and energy consumption has been tested.

As described in Chapter 3, during the learning process, the robot performs random arm movements and tries to estimate the position of its end-effector (the hand, tagged with a fiducial marker), analysing the frames grabbed from its head camera. A basic attentive mechanism has been implemented, composed by two modules: marker detection and motion detection. When a marker is detected, the head of the robot is rotated in order to focus on it, and the current configuration of the joint angles of the arm and of the neck are stored and coupled with the estimated 3D position of the hand. Due to the preliminary implementation presented in (Schillaci and Hafner, 2011b) and (Schillaci and Hafner, 2011a), the limited opening angle of the camera and the limited length of the robot's arms (like a child), for most of the time the robot has to rotate its head searching for the marker. The motion detection module is used in order to find the moving arm. Frame by frame, when the head is not moving, the optical flow between the current frame and the previous one is computed. The magnitude of the optical flow is fed into a CAMShift algorithm to find the centroid of the fastest moving area of the video to look at.

The results presented in (Schillaci and Hafner, 2011a) and (Schillaci and Hafner, 2011b) refer to three different types of movement strategies for motor babbling: Purely Random (PR), Random Walk (RW) and Inertial Random Walk (IRW). The babbling is performed on four degrees of freedom of the Nao's arm: shoulder pitch, shoulder roll, elbow yaw and elbow roll. In PR, random values are sampled from a uniform distribution over the range of each joint of the arm; in RW, random steps (increase/hold/decrease the joint by angle-step) are sampled from a uniform distribution; IRW is a random walk algorithm

Table A.1: Detection rates for the different random movement strategies.

	PR	RW	IRW
Detections per sec.	1.04	4.63	2.63
Max jump in deg.	665	40	40

with smoother movements, which simulates the inertia that a moving mass has when it changes the direction of the motion. Instant by instant, a random step is sampled from a uniform distribution, as in RW, and a small amount of the previous step is added to the current one, simulating the fact that the change of direction is not immediate, as the mass tends to follow the past movement by inertia.

Each strategy has been simulated for 8 minutes. Figure A.1 shows typical trajectories of the arm joints and of the neck joints for each type of random babbling strategy implemented in (Schillaci and Hafner, 2011a) and (Schillaci and Hafner, 2011b). PR generates sparse random commands in the action space; even if it can be thought as a good strategy able to explore uniformly the action space, the long jumps in the arm joints configuration very often increase the probability to lose the sight of the hand. This results in a very time consuming strategy with a low marker detection rate. Table A.1 shows some results for each strategy. Low detecting rates depend on a high probability that movements go outside the field of view of the camera, and on the time needed to catch again the arm by moving the head.

Table A.2: Energy Consumption Analysis

		PR	RW	IRW
Simulation	Distance Covered			
	PR	1.000	0.696	0.616
	RW	1.436	1.000	0.885
	IRW	1.622	1.130	1.000
Real Robot	Electric Current			
	PR	1.000	0.752	0.766
	RW	1.330	1.000	1.018
	IRW	1.306	0.982	1.000

Although IRW is the strategy that better resembles human motion, in the implementations presented in (Schillaci and Hafner, 2011a) RW seems to be the best strategy in terms of learning speed. IRW seems to perform worse than RW due to its tendency to follow the motion inertia towards areas wherein the hand is partially occluded by the shoulder

of the robot. The last row of Table A.1 represents, for each strategy, the maximum jump in degrees that a random movement can perform [1]

As an estimate of energy consumption, the sum of all the distances (in degrees) covered by each joint for each strategy during a certain amount of time has been also measured. These values have been also compared between the three strategies. In simulation, IRW resulted to be the cheapest strategy. In fact, when a joint is moving towards a certain position, if a new control command is generated towards the opposite direction of the current motion, in the simulated inertial strategy the direction is not changed instantaneously. Instead, the speed of the motion would be decreased, first, before changing direction. Changing the direction of motion instantaneously (as in the RW strategy) would consume more energy. Due to its fast changes of direction and movements, PR seems to be the worst strategy.

The sum of the distances is an estimate of energy consumption but, on the other hand, such a measurement would outcome the same amount of energy spent, for instance, for movements such as increasing a joint angle from 0 to 40 and increasing a joint angle from 0 to 20 and then decreasing it back to 0. For avoiding this, also the electric current applied to each servo has been measured and the averages of the total current applied to all the motor between the three strategies have been compared. In both energy measurements, also the two joints of the neck (which move accordingly to the attention system) have been considered.

The energy consumption analysis showed, this time, that RW is the best strategy. It seems that this is due to the fact that inertia is already intrinsic in the body of the robot and it does not need to be simulated in the real robot. Thus, IRW executed on a real robot requires more energy, because it accumulates a simulated inertia on the real one.

Both the results confirm that PR is the worst babbling strategy in learning a mapping between the joints configuration of the neck and that of the arm, because of the low marker detection rate and of the high energy dissipation. Analysing qualitatively the expectation of a human observer on the sensorimotor coordination skills of the robot, it can be noted that PR has also a significantly low rating. The robot is most of the time babbling and searching for the marker, due to the often long jump between an arm movement and the next one. RW and IRW have a higher rating.

[1] Ranges are (in degrees): ShoulderPitch: from -120 to 120; Shoulder Roll: from -95 to 0; Elbow Roll: from 0 to 90; Elbow Yaw: from -120 to 120. In RW and IRW, only a maximum step of 10 degrees is allowed for each joint.

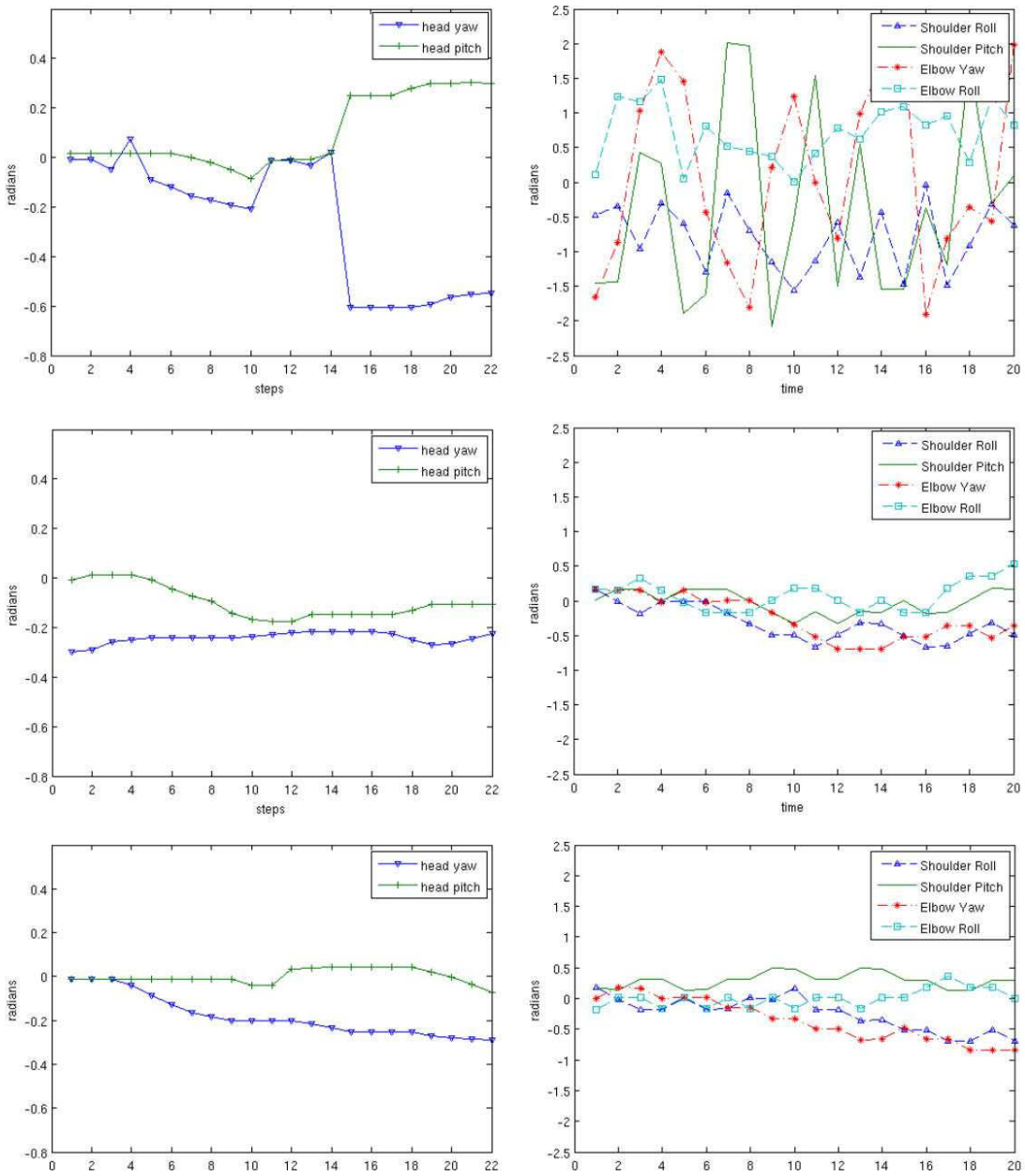


Figure A.1: Typical trajectories of the arm joints and of the neck joints for each type of random babbling strategy. In the left column of the figure, typical values of the joints angles of the neck for each strategy (PR, RW, IRW) are shown. The right column shows the values of the joint angles of the arm.

Appendix B

Evaluating the Effect of Saliency Detection and Attention Manipulation in Human-Robot Interaction

The ability to share the attention with another individual is essential for having intuitive interaction. Two relatively simple, but important prerequisites for this, saliency detection and attention manipulation by the robot, have been identified in (Bodiroža et al., 2011) and in (Schillaci et al., 2013a).

The rest of this appendix presents the statistical analysis reported in (Schillaci et al., 2013a), where we demonstrated that, by creating a saliency based attentional model combined with a robot ego-sphere and by adopting attention manipulation skills, the humanoid robot Aldebaran Nao can engage in an interaction with a human and start an interaction game including objects as a first step towards a joint attention.

Recently, interest has been focused on measuring the efficacy of robot behaviours and its perceived intelligence based on the evaluation from human users (Burghart and Steinfeld, 2008). Indeed, measuring human-robot interaction could suggest what and how to improve in the cognitive abilities and in the appearance of the robot.

In Chapter 3, saliency detection and attention manipulation skills implemented on the Aldebaran Nao robot have been introduced. In (Schillaci et al., 2013a), a partially preprogrammed motivation system has been implemented to show how different behaviours can result in the activation or deactivation of parts of the attention system, actually implementing a top-down approach for saliency detection, or in the activation of attention manipulation.

We tested our implementations in four combinations of activated parts of the attention system, which resulted in four different behaviours:

Exploration. In this state, the robot is attracted by movements, faces and objects, actually looking like exploring the surrounding environment.

Interaction. This behaviour reproduces the experiment done in (Hafner and Schillaci, 2011). The robot is looking and pointing at an object, if there is one.

Interaction avoidance. This behaviour implements the loss of interest and boredom. In this state the robot looks away from the marker.

Full interaction. This behaviour is composed as a sequence of the previous behaviours. The first performed action is *exploration*. Once the robot has detected a person to interact with and an object which can be used to draw the attention of the user, its motivation state changes to *interaction*, and after a certain period it switches to *interaction avoidance*, which is followed by *exploration*.

The experiments consisted of the robot performing the behaviours described before in four separate interaction sessions, one per each of the four behaviours. The experiment supervisor manually activated or deactivated them. The user sat in front of the robot at a distance of ca. 90 *cm*. For each person, each interaction test lasted one minute. We recorded the interaction with a standard camera (resolution 640×480) placed at ca. 2 meters perpendicularly to the robot-user axis. Beside the table where the robot was standing there was a scale drawn on a whiteboard for the visual estimation (estimated average error: $5cm$) of the distance between the nose of the user and the head of the robot and from the hand of the user and the head of the robot; according to the type of interaction, we noticed that the users move their hands closer to the robot.

In total 28 people participated in the survey, which results in a total of 112 questionnaires (four questionnaires per participant, one for each interaction). Some participants missed to answer some questions, but those were only a few questions. It is interesting to note that few participants had negative or neutral responses in all four experiments, regardless of the experiment, together with comments saying that Nao did not want anything because it is a machine. This might be perceived as a negative bias towards robots.

Of 28 participants, 8 were female (28.57 %) and 20 were male (71.43 %). There were 17 Germans, 2 Italians, 2 Serbians, 2 Poles, 1 Czech, 1 Dutch, 1 Estonian and 1 French. Regarding previous experience with robots, 25 persons (89.29 %) had none and 3 (10.71 %) had previous experience – one with industrial robots, one with Aldebaran Nao and one with Lego Mindstorms. The average age of the participants was 28.12 ($\sigma = 5.64$). Among the participants, 75 % had university level education and 25 % had high-school level education.

Unfortunately, not all the participants allowed us to film their interaction because of privacy reasons (even though we informed them that the data will be kept anonymous and videos will not be published against their will). The video database is composed of 10 videos for *exploration*, 7 for *interaction*, 8 for *avoid interaction* and 9 for *full interaction*.

Questionnaires

We conducted a qualitative, anonymous survey to evaluate how people perceive their interaction with the Nao. Questionnaires are often used to measure the user's attitude. Our first problem was related to what type of questionnaire to adopt. Developing a valid questionnaire can take a considerable amount of time and the absence of standardisation

makes it difficult to compare the results with other studies. That is why we decided to adopt standardised measurement tools for human-robot interaction, in addition to some metrics we found interesting for our research. We adopted as a part of our survey the Godspeed questionnaire (Bartneck et al., 2009) which uses semantic differential scales for evaluating the attitude towards the robot. Such a questionnaire contains questions (variables) about five concepts (latent variables): Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety (for a detailed description and for the set of questions, please refer to (Bartneck et al., 2009)).

Anthropomorphism refers to the attribution of human features and behaviours to non-human agents, such as animals, computers or robots. Anthropomorphism variables were (left value scored as 1, right value scored as 5): fake – natural, machinelike – humanlike, unconscious – conscious, artificial – lifelike, moving rigidly – moving elegantly.

Animacy is the property of alive agents. Robots can perform physical behaviours and reactions to stimuli. The participants' perception about robot animacy can give important insights for improving robot skills. Variables were: dead – alive, stagnant – lively, mechanical – organic, artificial – lifelike (different from the one in anthropomorphism, as related to the animacy), inert – interactive, apathetic – responsive.

Likeability may influence the user's judgments. Some studies indicate that people often make important judgments within seconds of meeting a person and it is assumed that people are able to judge also a robot (Bartneck et al., 2009). Likeability variables were: dislike – like, unfriendly – friendly, unkind – kind, unpleasant – pleasant, awful – nice.

Perceived Intelligence is one of the most important metrics for evaluating the efficacy of the implemented skills. It can depend on robot competence, but the duration of the interaction is also one of the most influencing factors, as users can become bored if the interaction is long and the vocabulary of the robot's behaviours is limited. Variables were: incompetent – competent, ignorant – knowledgeable, irresponsible – responsible, unintelligent – intelligent, foolish – sensible.

Perceived Safety is a metric for estimating the user's level of comfort when interacting with the robot and the perception of the level of danger. Variables were: anxious – relaxed, agitated – calm, quiescent – surprised (this variable was recoded, as explained in the next paragraph).

The reliability of the questionnaire was analysed by the authors of the Godspeed, who claim that such questions have sufficient internal consistency and reliability; to confirm this, we computed Cronbach's alpha[1] for each latent variable again. We found that Cronbach's alpha was negative ($\alpha = -1.111$) for the latent variable Perceived Safety, due to a negative average covariance among items. This violated reliability model assumptions for that set of variables, due to a miscoding of a variable. In fact, the questionnaire is written in such a way that high values of one variable mean the same thing as low values of the other variable; the miscoded variable was: Quiescent (scaled as 1) to Surprised (scaled as 5), probably due to the fact that participants intended quiescence as a syn-

[1]High Cronbach's alpha values are those greater than 0.5, which specify that the used set of variables are good for defining a certain concept.

onym for calmness (the previous variable was Agitated, coded as 1, or Calm, coded as 5). After recoding the quiescent – surprised variable, the Cronbach's alpha proved to be higher ($\alpha_{PerceivedSafety} = 0.839$)[2]. We did not found any other problems with the rest of the latent variables: $\alpha_{Anthropomorphism} = 0.825$, $\alpha_{Animacy} = 0.853$, $\alpha_{Likeability} = 0.813$, $\alpha_{PerceivedIntelligence} = 0.750$.

In addition to the Godspeed questionnaire, we introduced a new latent variable for measuring the concept of User Satisfaction, with two variables: frustrating – exciting and unsatisfying interaction – satisfying interaction (high Cronbach's alpha: $\alpha_{UserSatisfaction} = 0.799$).

Open questions were also introduced about the understanding of the behaviour of the robot, its desires, its aiming to interact or not, its successfulness, its gender (with the explanation of the chosen one), its age, type of communication during the interaction, expectations about future improvements and differences between Nao and humans.

Proxemics

According to the sociological concept of proxemics, humans, as well as animals, use to define personal spheres which delimit areas of physical distance that correlate reliably with how much people have in common (van Oosterhout and Visser, 2008). The boundaries of such spheres are determined by factors like gender, age and culture. Coming inside the sphere of another person may let him/her feel intimidated, or staying too far can be seen as cold or distant. Four spheres were identified, according to (van Oosterhout and Visser, 2008): Intimate Distance (from 0 to 45 cm), reserved for embracing, touching, whispering; Personal Distance (from 45 to 120 cm), reserved for friends; Social Distance (from 1.2 to 3.6 m), reserved for acquaintances and strangers; Public Distance (more than 3.6 m), reserved for public speaking.

However, in human-robot interaction, no assumptions about the existence of such boundaries have been made. The focus has been pointed on identifying those factors that influence interaction distance. Interaction distance can be influenced by factors like user age or gender, pet ownership, crowdedness in the environment or available space, as shown in (van Oosterhout and Visser, 2008) and (Takayama and Pantofaru, 2009). However, their analyses did not include users' perceptions about the behaviour or features of the robot.

We wanted to include proxemics measurement hoping to find some correlations between interaction distance and the factors treated in the questionnaire. We analysed participant behaviour also from measuring the distances between the face of the robot and the face of the user and between the face of the robot and the hand of the user[3].

As introduced in Section 3.4, proxemics analysis were done by gathering data from video recorded during the interaction sessions (Figure 3.15 shows a sample frame). The user sat in front of the robot at a distance of ca. 90 *cm*. We recorded the interaction with a standard camera (resolution 640×480) placed at ca. 2 meters perpendicularly to the

[2]For recoding, we intend flipping the variable: $1 = 5$, $2 = 4$, $3 = 3$, $4 = 2$, $5 = 1$.

[3]When interacting with the robot, participants did not use two hands at the same time. Almost all of them performed movements only with one arm, or at least they alternated between left and right. We registered only the movements from the active one.

axis robot-user. Beside the table where the robot was standing there was a scale drawn on a whiteboard for the visual estimation (estimated average error: $5cm$) of the distance between the nose of the user and the head of the robot and from the hand of the user and the head of the robot. Videos were annotated manually: every 5 seconds the face-face and face-hand distances were visually estimated by the operator, manually projecting their positions onto a scale drawn on the whiteboard.

Participants were sitting on a chair (they all started at the same distance to the robot), but they were told to feel free to interact in any way they considered more appropriate. However, it happened only in very few cases (only 2 participants) that they stood up. In both the cases, we gathered the face-face and face-hand distances as projected onto the horizontal line parallel to the table.

B.1 Results

This section presents the quantitative evaluation of the experiments reported in (Schillaci et al., 2013a).

In an earlier experiment we noticed some interesting patterns (see (Hafner and Schillaci, 2011) and (Bodiroža et al., 2011)). It seemed that if a person holds the object close to the robot's hand, then Nao's pointing will be perceived as a desire to grasp the object. This could indicate that, along with the hypothesis that pointing emerges from grasping, there is also a reverse connection – pointing can be perceived as grasping, if the object is too close to the hand[4]. Furthermore, most of the participants in the preliminary experiment responded that Nao was either likeable or very likeable and that the speed of experiment was good (out of three possible answers: too fast, good and too slow), even though the execution speed was lower than in the current experiment. All participants in the preliminary experiment, except one, had no previous experience with robots.

Figure B.1 shows the means and the standard deviations of the responses.

First, we checked whether the distributions of the collected data are normal or not, in order to select the proper statistical tests. For each variable (that is, for each question), we looked at the superimposition of the histogram of the data with a normal curve characterised by the mean and the variance of the data. Almost all the histograms did not fit well together with the corresponding normal curves. Thus, we checked the kurtosis and the skewness of the data[5], in order to have a more precise measurement of the normality of the distributions. The distributions of all the variables related to the questionnaire had kurtosis and skewness between -2 and $+2$, while 17 out of 64 distributions related to the variables of the proxemics analysis[6] did not.

[4]The robot platform we used has no movable fingers and it is unable to grasp an object.

[5]In general, when kurtosis and skewness are between -2 and $+2$, the data is not too far away from a normal distribution. When that is not the case, corrections (like Box-Cox transformations) can be applied to the data in order to apply the tests that have assumptions of normality.

[6]For each of the four behaviours performed by the robot, we created two variables for the average value and variance of the distance between the face of the Nao and the nose of the participant for the following cases: during the first 15 seconds of the interaction, between the 15th second and the 45th second of the interaction, and during the last 15 seconds of the interaction (in total 6 variables). The same variables

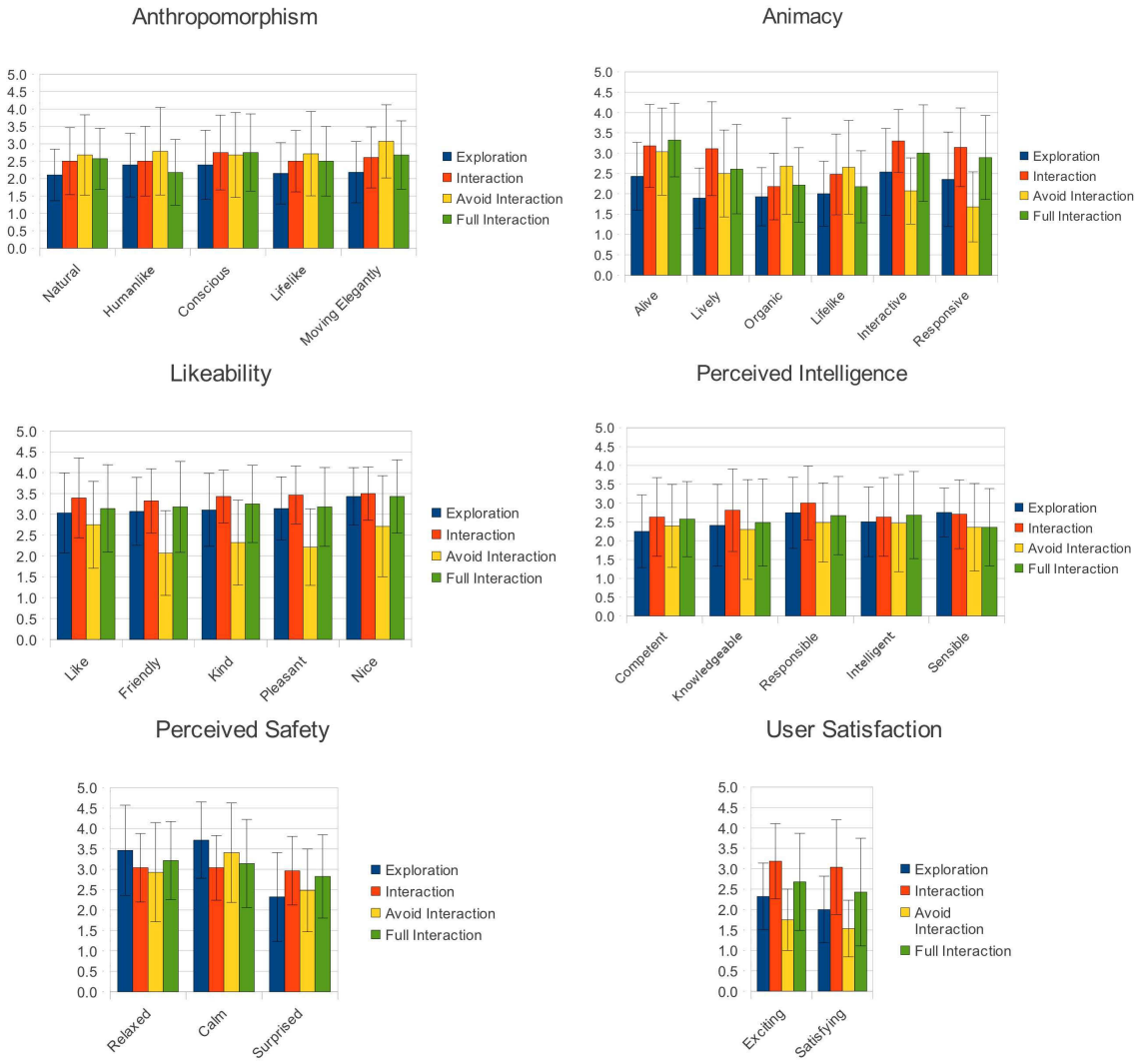


Figure B.1: These graphs show the results taken from the Godspeed questionnaire.

Due to the non-normality of such distributions, it seems to be more appropriate applying non-parametric statistical tests for the whole analysis. However, the use of ANOVA on Likert-scale data and without the assumption of normality of the distributions of the data to be analysed is controversial. In general, researchers claim that only non-parametric statistics should be used on Likert-scale data and when the normality assumption is violated. (Vallejo et al., 2010), instead, found that the Repeated Measures ANOVA[7] was robust toward the violation of normality assumption. Simulation results of (Schmider et al., 2010) confirm also this observation, since they found in their Monte Carlo study that the empirical Type I and Type II errors in ANOVA were not affected by the violation of assumptions.

Correlations

A Spearman's Rank Order correlation[8] was run to determine the relationship between perceived factors and between them and average human-robot distances. Each run was done for each experimental session (*exploration*, *interaction*, *interaction avoidance* and *full interaction*).

Tables B.1 and B.2 show some of the most relevant correlations. In addition to the data shown in the tables, it has to be noted that in the *exploration* test there was a strong, positive correlation between almost all the anthropomorphism variables and the perceived intelligence attributes related to competence and knowledge; in *interaction*, the higher the likeability of the robot, the higher the variance of face-face distance during all the interaction tests ($r = 0.805$, $P = 0.029$, $N = 7$); in *full interaction*, perceived intelligence was found to be positively correlated with almost all the other variables (except those related to perceived safety) with $r > 0.5$ and almost always significant at the 0.01 level.

Repeated Measures ANOVA

Because the participants of the four different observations were the same in each group, we adopted the Repeated measures ANOVA test (post-hoc test using Bonferroni correction) for the analysis of variances. Also known as within-subjects ANOVA test, repeated measures ANOVA is the equivalent of the one-way ANOVA but for related, not independent groups. We performed the test on all the dependent variables[9] .

---

were created for analysing the distance between the face of the Nao and the user's hand closest to the robot.

[7]Repeated measures ANOVA compare the average score for a single group of subjects at multiple time periods (observations).

[8]Spearman's correlation coefficient is non-parametric, looks at ranked (coded) variables (without looking at the data directly) and does not have the normality assumption on the distributions, thus it can be used for skewed or ordinal variables. We ran the correlation with 2-tailed test of significance. Missing values were excluded with cases pairwise.

[9]Mauchly's test has been used as statistical test for validating repeated measures ANOVA. It tests the sphericity, which is related to the equality of the variances of the differences between levels of the repeated measures factor. Sphericity, an assumption of repeated measures ANOVA, requires that the variances for each set of difference scores are equal. Sphericity can not be assumed when the significance level of the Mauchly's test is < 0.05. Violations of sphericity assumption can invalidate the analysis

Post-hoc tests revealed that the four different behaviours performed by the robot have not changed significantly the participants' perception of the anthropomorphic attributes related to naturalness, humanlikeness, consciousness and artificiality. Table B.3 shows the statistically significant results of repeated measures ANOVA on the questionnaire variables.

Proxemics variables contain a high number of missing values. In order to perform repeated measures ANOVA on those variables, we had to replace missing values with multiple imputation ($n = 20$). New samples were created, where proxemics information was inferred using the questionnaire variables as predictors[10].

Table B.4 shows the statistically significant results of the repeated measures ANOVA on the proxemics variables.

Latent Growth Curve Model

A latent growth curve model was also used to assess the change in user perception over the four behaviours. This model uses a structural equation to estimate two latent variables, the slope and intercept, to assess the average linear change across the measurements, where the individual measurements are the indicators of the latents[11]. The estimated population distribution of the linear change (or growth) trajectory, denoted by the slope and the intercept of a linear function, are derived from this structural equation model. The estimator selected for the procedure was a Bayesian estimator with non-informative priors[12]. All calculations were produced with Mplus 6.11.

The estimated slopes for many of the items were almost all positive, with also positive credibility intervals, meaning that there is a significant positive trend in the average score from the first observation (*exploration*) to the last observation (*full interaction*)[13].

---

conclusions, but corrections can be applied to alter the degrees of freedom in order to produce a more accurate significance value, like the Greenhouse-Geisser correction. When the significance level of the Greenhouse-Geisser estimate is < 0.05, statistical significant differences revealed by post-hoc test can be elicited from the pairwise comparisons between the observations. Repeated Measures ANOVA does not tell where the differences between groups lie. When repeated measures ANOVA is statistically significant (both with sphericity assumption not violated or with Greenhouse-Geisser correction), post-hoc tests with multiple comparisons can highlight exactly where these differences occur.

[10]For multiple imputation, all the available variables that can predict the values of missing data should be included.

[11]The loadings are constrained to be 1 for the intercept latent and to 0 to 3 (depending of the time of measurement) for the slope latent.

[12]This estimation strategy was appropriate as the more commonly used maximum likelihood estimator often produces biased (or often inestimable) results with such small sample sizes. The Bayesian estimator is more robust to both small samples and violation of distributional assumptions that could emerge from small samples.

[13]Further analysis can be done on piecewise linear growth, for breaking up the curvilinear growth trajectories into separate linear components, thus for analysing whether there was an increase or a decrease between *exploration* and *interaction*, between (*interaction* to *interaction avoidance*, and so on.

Table B.1: Most relevant correlations (Part 1). For having the full tables, please ask the authors.

		Exploration			Interaction			Inter. Avoidance			Full Interaction		
Variables correlated		R	p	N	R	p	N	R	p	N	R	p	N
Anthropomor.: humanlike	Animacy: alive	0.581	0.001	28	0.654	0.000	28	0.665	0.000	28	0.546	0.003	28
Anthropomor.: humanlike	Animacy: interactive	0.513	0.005	28	0.605	0.001	27	Stat. not signif.			0.504	0.006	28
Anthropomor.: humanlike	Perc. Intel.: knowledgeable	0.416	0.031	28	0.562	0.003	26	0.571	0.002	27	0.606	0.001	27
Anthropomor.: humanlike	Perc. Intel.: competent	0.476	0.011	28	0.677	0.000	27	0.623	0.000	28	0.564	0.002	28
Anthropomor.: humanlike	Perc. Intel.: intelligent	Stat. not signif.			0.559	0.002	27	0.573	0.001	28	**0.713**	0.000	28
Anthropomor.: natural	Perc. Intel.: knowledgeable	0.498	0.008	27	0.557	0.003	26	Stat. not signif.			0.553	0.003	27
Anthropomor.: natural	Perc. Intel.: competent	0.565	0.002	28	0.612	0.001	27	Stat. not signif.			0.572	0.001	28
Anthropomor.: moving elegantly	Perc. Intel.: knowledgeable	0.697	0.000	26	0.399	0.044	26	0.654	0.000	27	0.422	0.028	27
Anthropomor.: moving elegantly	Perc. Intel.: competent	**0.694**	0.000	27	0.483	0.011	27	0.542	0.003	28	0.454	0.015	28
Anthropomor.: moving elegantly	Likeability: friendly	Stat. not signif.			Stat. not signif.			-0.500	0.007	28	Stat. not signif.		
Anthropomor.: lifelike	Variance Face-Hand (15"-45")	Stat. not signif.			Stat. not signif.			Stat. not signif.			**-0.786**	0.012	9
Animacy: lifelike	Likeability: friendly	Stat. not signif.			0.663	0.001	23	-0.425	0.043	23	Stat. not signif.		
Animacy: interactive	Anthropomor.: lifelike	**0.673**	0.012	13	0.660	0.000	27	Stat. not signif.			0.556	0.002	28
Animacy: interactive	Likeability: friendly	0.398	0.036	28	0.451	0.018	27	Stat. not signif.			0.655	0.000	28
Animacy: interactive	Perc. Intel.: intelligent	**0.462**	0.013	28	**0.705**	0.000	26	**0.403**	0.033	28	**0.619**	0.000	28
Animacy: interactive	User Satisf.: exciting	**0.710**	0.000	28	0.551	0.004	26	Stat. not signif.			**0.706**	0.000	28
Animacy: interactive	User Satisf.: satisfying	0.470	0.012	28	0.687	0.000	26	Stat. not signif.			0.725	0.000	28
Animacy: responsive	Average Face-Face distance (60s.)	0.633	0.037	11	Stat. not signif.			Stat. not signif.			Stat. not signif.		
User Satisf.: satisfying interaction	Anthropomor.: moving elegantly	0.505	0.007	27	0.482	0.011	27	Stat. not signif.			0.390	0.040	28
User Satisf.: satisfying interaction	Anthropomor.: lifelike	0.576	0.002	26	**0.653**	0.000	27	Stat. not signif.			0.516	0.005	28
User Satisf.: satisfying interaction	Animacy: responsive	**0.696**	0.000	28	**0.722**	0.000	27	Stat. not signif.			**0.673**	0.000	28
Perceived Safety: quiescent	Average Face-Face dist. (last 15")	Stat. not signif.			Stat. not signif.			-0.879	0.009	7	Stat. not signif.		
Perceived Safety: quiescent	Average Face-Hand dist. (last 15")	Stat. not signif.			Stat. not signif.			-0.805	0.029	7	Stat. not signif.		

Table B.2: Most relevant correlations (Part 2). For having the full tables, please ask the authors.

Variables correlated		Exploration			Interaction			Inter. Avoidance			Full Interaction		
		R	p	N	R	p	N	R	p	N	R	p	N
Variance Face-Hand dist. (0”-15”)	Perceived Safety: quiescent	Stat. not signif.			Stat. not signif.			Stat. not signif.			-0.670	0.048	9
Variance Face-Hand dist. (60s.)	Likeability: friendly	Stat. not signif.			0.802	0.030	7	Stat. not signif.			-0.673	0.047	9
Variance Face-Hand dist. (60s.)	Likeability: kind	Stat. not signif.			Stat. not signif.			Stat. not signif.			**-0.738**	0.023	9
Variance Face-Hand dist. (60s.)	Likeability: pleasant	Stat. not signif.			Stat. not signif.			Stat. not signif.			**-0.829**	0.006	9
Variance Face-Hand dist. (60s.)	User Satisf.: satisfying interaction	Stat. not signif.			Stat. not signif.			Stat. not signif.			**-0.738**	0.023	9
Variance Face-Face dist. (15”-45”)	Likeability: friendly	Stat. not signif.			0.809	0.028	7	Stat. not signif.			Stat. not signif.		
Average Face-Face dist. (60s.)	User Satisf.: exciting	Stat. not signif.			Stat. not signif.			Stat. not signif.			0.709	0.032	9
Average Face-Face dist. (60s.)	Perc. Intellingence: intelligent	Stat. not signif.			Stat. not signif.			Stat. not signif.			0.729	0.026	9

Table B.3: Statistically significant results of repeated measures ANOVA on the questionnaire variables. Cases with sphericity assumption violated were corrected with Greenhouse-Geisser method. The table shows the statistically significant pairwise comparisons (illustrating the changes in means from an observation to another), taken from the post-hoc test with Bonferroni correction.

Variable	Sphericity assumed	From observ.	To observ.	Mean difference	Std. error	Significance
Anthr.: moving elegantly	no	1	3	-0.889	0.252	0.010
		2	3	-0.481	0.154	0.026
Animacy: alive	yes	1	2	-0.75	0.203	0.006
		1	4	-0.893	0.165	0.000
Animacy: lively	yes	1	2	-1.222	0.284	0.001
		1	4	-0.741	0.224	0.016
Animacy: organic	no	1	2	-0.750	0.239	0.025
		2	3	-0.500	0.159	0.024
Animacy: interactive	yes	1	2	-0.815	0.251	0.019
		2	3	1.222	0.202	0.000
		3	4	-1.000	0.233	0.001
Animacy: responsive	yes	1	2	-0.786	0.259	0.032
		2	3	1.464	0.260	0.000
		3	4	-1.214	0.243	0.000
Likeability: friendly	no	1	3	1.000	0.230	0.001
		2	3	1.250	0.270	0.001
		3	4	-1.107	0.274	0.002
Likeability: kind	yes	1	3	0.786	0.243	0.019
		2	3	1.107	0.248	0.001
		3	4	-0.929	0.224	0.002
Likeability: pleasant	no	1	3	0.929	0.185	0.000
		2	3	1.250	0.222	0.000
		3	4	-0.964	0.238	0.002
Likeability: nice	no	1	3	0.714	0.198	0.008
		2	3	0.786	0.249	0.023
Perceived Safety: quiescent	yes	1	2	0.593	0.194	0.031
		2	3	-0.481	0.154	0.026
User satisfaction: exciting	no	1	2	-0.852	0.218	0.004
		2	3	1.407	0.234	0.000
		2	4	0.444	0.154	0.047
		3	4	-0.963	0.285	0.014
User satisfaction: satisfying	yes	1	2	-1.037	0.196	0.000
		2	3	1.519	0.222	0.000
		3	4	-0.963	0.229	0.002

Table B.4: Statistically significant results of repeated measures ANOVA on the proxemics variables. Cases with sphericity assumption violated were corrected with Greenhouse-Geisser method. The table shows the statistically significant pairwise comparisons (illustrating the changes in means from an observation to another), taken from the post-hoc test with Bonferroni correction. Missing values were replaced with multiple imputations. The new dataset contained 560 samples. Abbreviations: AV: average; VAR: variance; FF: distance between the face of the robot and the face of the user; FH: distance between the face of the robot and the closest hand of the user; all: considering the whole duration of the test (60 seconds).

Variable	Sphericity assumed	From observ.	To observ.	Mean difference	Std. error	Significance
AV FF all	no	1	2	**13.675**	0.419	0.000
		1	3	11.876	0.302	0.000
		1	4	9.734	0.355	0.000
		2	3	**-1.799**	0.360	0.000
		2	4	**-3.941**	0.436	0.000
		3	4	-2.142	0.280	0.000
VAR FF all	no	1	2	-6.218	2.058	0.016
		1	3	-82.552	3.014	0.000
		1	4	**14.558**	1.914	0.000
		2	3	-76.334	2.361	0.000
		2	4	20.776	1.633	0.000
		3	4	97.110	2.862	0.000
AV FH all	no	1	2	14.767	0.544	0.000
		1	3	**18.439**	0.547	0,000
		1	4	**21.423**	0.539	0.000
		2	3	3.672	0.294	0.000
		2	4	6.656	0.314	0.000
		3	4	2.985	0.307	0.000
VAR FH all	no	1	2	**20.337**	3.861	0.000
		1	3	**-217.131**	7.246	0.000
		2	3	**-237.468**	6.718	0.000
		2	4	-28.076	5.602	0.000
		3	4	**209.392**	6.904	0.000

Bibliography

Abend, W., Bizzi, E., Morasso, P., et al. (1982). Human arm trajectory formation. *Brain: a journal of neurology*, 105(Pt 2):331.

Akgun, B. and Tunaoglu, D.and Sahin, E. (2010). Action recognition through an action generation mechanism. In *International Conference on Epigenetic Robotics (EPIROB)*.

Allen, J. G., Xu, R. Y. D., and Jin, J. S. (2004). Object tracking using camshift algorithm and multiple quantized feature spaces. In *Proceedings of the Pan-Sydney area workshop on Visual information processing*, VIP '05, pages 3–7, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34.

Atkeson, C. G. (1989). Learning arm kinematics and dynamics. *Annual Review of Neuroscience*, 12(1):157–183. PMID: 2648948.

Baranes, A. and Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.*, 61(1):49–73.

Baron-Cohen, S. (1995). *Mindblindness. An essay on autism and theory of mind*. MIT Press.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59(1):617–645.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81.

Berthouze, L. and Metta, G. (2005). Editorial: Epigenetic robotics: modelling cognitive development in robotic systems. *Cogn. Syst. Res.*, 6(3):189–192.

Billard, A., Calinon, S., Dillmann, R., and Schaal, S. (2008). Survey: Robot programming by demonstration. Handbook of Robotics, . chapter 59, 2008.

Blake, R. and Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.*, 58:47–73.

Blakemore, S. J., Goodbody, S. J., and Wolpert, D. M. (1998a). Predicting the consequences of our own actions: The role of sensorimotor context estimation. *Journal of Neuroscience*, 18:7511–7518.

Blakemore, S. J., Wolpert, D., and Frith, C. (2000). Why can't you tickle yourself? *NeuroReport*, 11(11):11–16.

Blakemore, S. J., Wolpert, D. M., and Frith, C. D. (1998b). Central cancellation of self-produced tickle sensation. *Nat Neurosci*, 1(7):635–640.

Bodiroža, S., Schillaci, G., and Hafner, V. (2011). Robot ego-sphere: An approach for saliency detection and attention manipulation in humanoid robots for intuitive interaction. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 689–694.

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, 314(5802):1118–1121.

Bosbach, S., Cole, J., Prinz, W., and Knoblich, G. (2005). Inferring another's expectation from action: the role of peripheral sensation. *Nature Neuroscience*, 8:1295–1297.

Bril, B., Rein, R., Nonaka, T., Wenban-Smith, F., and Dietrich, G. (2010). The role of expertise in tool use: Skill differences in functional action adaptations to task constraints. *Journal of Exp. Psyc.: Human Perception and Performance*, 36(4):825 – 839.

Burghart, C. R. and Steinfeld, A., editors (2008). *Proceedings of the Metrics for Human-Robot Interaction Workshop at the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008)*.

Butterworth, G. (1995). Origins of mind in perception and action. In Moore, C. and Dunham, P., editors, *Joint attention: its origins and role in development*, pages 29–40. Lawrence Erlbaum Associates.

Butterworth, G. and Hopkins, B. (1988). Hand-mouth coordination in the new-born baby. *British Journal of Developmental Psychology*, 6(4):303–314.

Casile, A. and Giese, M. A. (2006). Nonvisual motor training influences biological motion perception. *Current Biology*, 16(1):69 – 74.

Cooper, L. and Podgorny, P. (1975). *Demonstration of a Mental Analog of an External Rotation*. CHIP report. Center for Human Information Processing, University of California, San Diego.

Dautenhahn, K. (2003). Roles and functions of robots in human society: implications from research in autism therapy. *Robotica*, 21(4):443–452.

Dautenhahn, K. and Billard, A. (1999). Studying robot social cognition within a developmental psychology framework. In *In Proceedings of the 3rd International Workshop on Advanced Mobile Robots*.

Dearden, A. (2008). *Developmental learning of internal models for robotics*. PhD thesis, Imperial College London.

Dearden, A. and Demiris., Y. (2005). Learning forward models for robots. In *Int. Joint Conferences on Artificial Intelligence*, page 1440, Edinburgh.

Demiris, Y. and Dearden, A. (2005). From motor babbling to hierarchical learning by imitation: A robot developmental pathway. In *In EpiRob*.

Demiris, Y. and Khadhouri, B. (2006). Hierarchical attentive multiple models for execution and recognition of actions. *Robotics and Autonomous Systems*, 54(5):361–369.

de'Sperati, C. and Viviani, P. (1997). The relationship between curvature and velocity in two-dimensional smooth pursuit eye movements. *The Journal of Neuroscience*, 17(10):3932–3945.

Desrochers, S., Morisette, P., and Ricard, M. (1995). Two perspectives on pointing in infancy. In Moore, C. and Dunham, P., editors, *Joint Attention: its origins and role in development*, pages 85–101. Lawrence Erlbaum Associates.

Di Nuovo, A. G., Marocco, D., Di Nuovo, S., and Cangelosi, A. (2013). Autonomous learning in humanoid robotics through mental imagery. *Neural Networks*, 41:147–155.

Dreyfus, H. L. (1972). *What computers can't do: a critique of artificial reason,*. Harper & Row, New York.

Eskandar, E. N. and Assad, J. A. (1999). Dissociation of visual, motor and predictive signals in parietal cortex during visual guidance. *Nature Neuroscience*, 2:88–93.

Eskenazi, T., Grosjean, M., Humphreys, G., and Knoblich, G. (2009). The role of motor simulation in action perception: a neuropsychological case study. *Psychological Research PRPF*, 73(4):477–485.

Fleming, K. A., Peters, R. A., and Bodenheimer, R. E. (2006). Image mapping and visual attention on a sensory ego-sphere. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 241–246.

Frak, V., Paulignan, Y., and Jeannerod, M. (2001). Orientation of the opposition axis in mentally simulated grasping. *Exp. Brain Res.*, 136:120–127.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia / Christopher D. Frith*. L. Erlbaum Associates Hove, U.K. ; Hillsdale, U.S.

Gallese, V. (2007). Before and below theory of mind: embodied simulation and the neural correlates of social cognition. *Phil. Trans. of the Royal Society B, 362(1480):659-669, Apr. 2007.*, 362(1480):659–669.

Gentsch, A. and Schütz-Bosbach, S. (2011). I did it: Unconscious expectation of sensory consequences modulates the experience of self-agency and its functional signature. *J. Cognitive Neuroscience*, 23(12):3817–3828.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin Company.

Gibson, J. J. (1977). *The Theory of Affordances*. Lawrence Erlbaum.

Goodbody, S. J., Husain, M., and Wolpert, D. M. (1998). Maintaining internal representations: the role of the human superior parietal lobe. *Nature Neuroscience*, 1:529–533.

Grezes, J., Armony, J. L., Rowe, J. B., and Passingham, R. E. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *Neuroimage*, 18(4):928–937.

Hafner, V. V. and Schillaci, G. (2011). From field of view to field of reach - could pointing emerge from the development of grasping? *Frontiers in Computational Neuroscience*, (17).

Haruno, M., Wolpert, D. M., and Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, 13:2201–2220.

Hoffmann, H. (2007). Perception through visuomotor anticipation in a mobile robot. *Neural Networks*, 20:22–33.

Hoffmann, H. and Möller, R. (2004). Action selection and mental transformation based on a chain of forward models. In *Proc. of the 8th Int. Conference on the Simulation of Adaptive Behavior*, pages 213–222, Cambridge, MA. MIT Press.

Holt, R. R. (1964). Imagery: The return of the ostracized.

Ijspeert, A. J., Nakanishi, J., and Schaal, S. (2001). Trajectory formation for imitation with nonlinear dynamical systems. In IEEE, editor, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259.

Jansen, B., Boer, B., and Belpaeme, T. (2004). You did it on purpose! towards intentional embodied agents. In Iida, F., Pfeifer, R., Steels, L., and Kuniyoshi, Y., editors, *Embodied Artificial Intelligence*, volume 3139 of *Lecture Notes in Computer Science*, pages 271–277. Springer Berlin Heidelberg.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, pages 103–109.

Jevtic, A., Lucet, E., Kozlov, A., and Gancet, J. (2012). Intro: A multidisciplinary approach to intelligent human-robot interaction. In *World Automation Congress (WAC), 2012*, pages 1–6.

Jonsson, G. K. and Thorisson, K. R. (2010). Evaluating multimodal human-robot interaction: A case study of an early humanoid prototype. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*, MB '10, pages 9:1–9:4. ACM.

Jordan, M. I. and Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *cognitive science*, 16:307–354.

Kaplan, F. and Hafner, V. V. (2004). The challenges of joint attention. *Interaction Studies*, 7:67–74.

Kiverstein, J. (2007). Could a robot have a subjective point of view? *Journal of Consciousness Studies*, 14(7):127–139.

Knoblich, G. and Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, 12(6):467–472.

Kosslyn, S. M. (1980). *Image and Mind*. Harvard University Press, Cambridge, MA.

Kosslyn, S. M., Reiser, B. J., and Ball, T. M. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, pages 47–60.

Kozlov, A., Gancet, J., Letier, P., Schillaci, G., Hafner, V. V., Fonooni, B., Hellstrom, T., Nevatia, Y., and Govindaraj, S. (2013). Toward a search and rescue field robotic assistant. In *Proceedings of the 11th IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR2013), Linkping, Sweden.*

Kuhl, P. K. and Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, 100(4 Pt 1):2425–2438.

Lacquaniti, F., Terzuolo, C., and Viviani, P. (1983). The law relating the kinematic and figural aspects of drawing movements. *Acta psychologica*, 54(1):115–130.

Laird, J. D. (1974). Self-attribution of emotion: The effects of expressive behavior on the quality of emotional experience. *Journal of Personality and Social Psychology*, 29:475–486.

Lara, B. and Rendon, J. M. (2006). Prediction of undesired situations based on multimodal representations. In *Proceedings of the Electronics, Robotics and Automotive Mechanics Conference - Volume 01*, pages 131–136, Washington, DC, USA. IEEE Computer Society.

Lara, B., Rendon, J. M., and Capistran, M. (2007). Prediction of multi-modal sensory situations, a forward model approach. In *Proc. of the 4th IEEE Latin America Robotics Symposium*, volume 1.

Leslie, A. M. (1994). Tomm, toby, and agency: Core architecture and domain specificity. *In L. A. Hirschfeld and S. A. Gelman (Eds.), Mapping the mind: Domain specificity in cognition and culture. Cambridge. Cambridge University Press.*

Lindblom, J. and Ziemke, T. (2002). Social situatedness: Vygotsky and beyond. 94:71–78.

Loren, L. A. and Dietrich, E. (1997). Merleau-ponty, embodied cognition, and the problem of intentionality. *Cybernetics and Systems*, 28:345–58.

Loula, F., Prasad, S., Harber, K., and Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):210.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, 15:151–190.

Maravita, A. and Iriki, A. (2004). Tools for the body (schema). *Trends Cogn Sci*, 8(2):79–86.

Maravita, A., Spence, C., and Driver, J. (2003). Multisensory integration and the body schema: close to hand and within reach. *Current Biology*, 13:531–539.

Maturana, H. and Varela, F. (1987). *The tree of knowledge: the biological roots of human understanding.* New Science Library. Shambhala.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence (august 31, 1955). *AI Magazine, Association for the Advancement of Artificial Intelligence*, 27(4).

McCarty, M. E., Clifton, R. K., Ashmead, D. H., Lee, P., and Goubet, N. (2001). Biobehaviorial Development, Perception, and Action: How Infants Use Vision for Grasping Objects. *Child Development*, 72:973–987.

Megill, J. L. (2003). Conceptualized direct perception: A hybrid theory of vision. In *Master Thesis.* Louisiana State University, Department of Philosophy and Religious Studies.

Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198(4312):75–78.

Meltzoff, A. N. and Moore, M. K. (1997). Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6:179–192.

Metta, G. (2000). Babyrobot – a study on sensori-motor development.

Metta, G., Sandini, G., Natale, L., Craighero, L., and Fadiga, L. (2006). Understanding mirror neurons: A bio-robotic approach. *Interaction Studies*, 7(2):197–232.

Miall, R. C. and Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, 9(8):1265–1279.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends Cogn Sci*, 7(3):141–144.

Mingers, J. (2001). Embodying information systems: the contribution of phenomenology. *Information and Organization*, 11(2):103–128.

Möller, R. and Schenck, W. (2008). Bootstrapping cognition from behavior–a computerized thought experiment. *Cognitive Science*, 32(3):504–542.

Moore, C., Angelopoulos, M., and Bennett, P. (1997). The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1):83 – 92.

Morse, A. F., Greef, J. D., Belpaeme, T., and Cangelosi, A. (2010). Epigenetic robotics architecture (ERA). *IEEE Transactions on Automnomous Mental Development*, 2(4):325–339.

Morton, J. and Johnson, M. H. (1991). Conspec and conlern: A two-process theory of infant face recognition. *Psychological Review*, 98:164–181.

Nagy, E. and Molnar, P. (2004). Homo imitans or homo provocans? Human imprinting model of neonatal imitation. *Infant Behavior and Development*, 27(1):54–63.

Noriega, L. (2005). Multilayer perceptron tutorial. *Staffordshire University*.

O'Keefe, J. and Recce, M. L. (1993). Phase relationship between hippocampal place units and the eeg theta rhythm. hippocampus 3:317330. *Hippocampus*, 3:317–330.

Olsson, L., Nehaniv, C. L., and Polani, D. (2006). From unknown sensors and actuators to actions grounded in sensorimotor perceptions. *Connection Science*, 18:2006.

Peters, R. A., Hambuchen, K. E., Kawamura, K., and Wilkes, D. M. (2001). The sensory ego-sphere as a short-term memory for humanoids. In *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, pages 451–460.

Pfeifer, R. and Bongard, J. (2006). How the body shapes the way we think: A new view of intelligence. *The MIT Press*.

Pfeifer, R. and Gómez, G. (2009). Creating brain-like intelligence. chapter Morphological Computation — Connecting Brain, Body, and Environment, pages 66–83. Springer-Verlag, Berlin, Heidelberg.

Piaget, J. (1983). Piaget's Theory. *Handbook of Child Psychology*.

Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, 2(3):211–228.

Pylyshyn, Z. (1977). What the minds eye tells the minds brain: A critique of mental imagery. In Nicholas, J., editor, *Images, Perception, and Knowledge*, volume 8 of *The University of Western Ontario Series in Philosophy of Science*, pages 1–36. Springer Netherlands.

Reissland, N. (1988). Neonatal imitation in the first hour of life: Observations in rural nepal. *Developmental Psychology*, 24(4):464–469.

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27(1):169–192. PMID: 15217330.

Rizzolatti, G., Fadiga, L., Fogassi, L., and Gallese, V. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141.

Robinson, H. (2012). Dualism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Rochat, P. (1993). Chapter 10 hand-mouth coordination in the newborn: Morphology, determinants, and early development of a basic act. In Savelsbergh, G. J., editor, *The Development of Coordination in Infancy*, volume 97 of *Advances in Psychology*, pages 265 – 288. North-Holland.

Rochat, P. (1998). Self-perception and action in infancy. *Experimental Brain Research*, 123(1-2):102–109.

Rochat, P. and Morgan, R. (1998). Two functional orientations of self-exploration in infancy. *British Journal of Developmental Psychology*, 16(2):139–154.

Rolf, M., Steil, J., and Gienger, M. (2011). Online goal babbling for rapid bootstrapping of inverse models in high dimensions. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–8.

Rolf, M. and Steil, J. J. (2012). Goal babbling: a new concept for early sensorimotor exploration. Proceedings of Workshop on Developmental Robotics. IEEE.

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *ICRA 2008. IEEE International Conference on Robotics and Automation*, pages 962–967.

Saegusa, R., Metta, G., Sandini, G., and Sakka, S. (2009). Active motor babbling for sensorimotor learning. In *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference on*, pages 794–799.

Scassellati, B. (2001). Foundations for a theory of mind for a humanoid robot.

Schillaci, G., Bodiroa, S., and Hafner, V. (2013a). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, 5(1):139–152.

Schillaci, G., Hafner, V., and Lara, B. (2013b). I know how i would do it. internal simulations of sensorimotor experience. In *Robotics Science and Systems Conference, Proceedings of the Workshop in Active learning in robotics: Exploration, Curiosity, and Interaction. Berlin, Germany*.

Schillaci, G., Hafner, V., Lara, B., and Grosjean, M. (2013c). Is that me? sensorimotor learning and self-other distinction in robotics. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 223–224.

Schillaci, G. and Hafner, V. V. (2011a). Prerequisites for intuitive interaction - on the example of humanoid motor babbling. In *Proc. of the Workshop on: The role of expectations in intuitive human-robot interaction (at HRI 2011)*, pages 23–27.

Schillaci, G. and Hafner, V. V. (2011b). Random Movement Strategies in Self-Exploration for a Humanoid Robot. In *Proc. of the Intern. Conf. on Human-Robot Interaction 2011*, pages 245–246.

Schillaci, G., Hafner, V. V., and Lara, B. (2012a). Coupled inverse-forward models for action execution leading to tool-use in a humanoid robot. In *In Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, Boston.

Schillaci, G., Lara, B., and Hafner, V. (2012b). Internal simulations for behaviour selection and recognition. In Salah, A., Ruiz-del Solar, J., Meriçli, Ç., and Oudeyer, P.-Y., editors, *Human Behavior Understanding*, volume 7559 of *Lecture Notes in Computer Science*, pages 148–160. Springer Berlin / Heidelberg.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., and Bhner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, 6(4):147–151.

Sheldon, M. T. (2012). *Intrinsically Motivated Developmental Learning of Communication in Robotic Agents*. PhD thesis, Aberystwyth University.

Shepard, R. N. and Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, 171(3972):701–703.

Sima, J. and Freksa, C. (2012). Towards computational cognitive modeling of mental imagery. *KI - Kuenstliche Intelligenz*, 26(3):261–267.

Smith, D. W. (2011). Phenomenology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Spivey, M. J., Tyler, M. J., Richardson, D. C., and Young, E. E. (2000). Eye movements during comprehension of spoken scene descriptions. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society (pp. 487492). Mahwah, NJ: Lawrence Erlbaum Associates Inc.*

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, HRI '06, pages 33–40. ACM.

Strack, F., Martin, L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54.

Suder, K. and Wörgötter, F. (2000). The control of low level information flow in the visual system. *Rev. Neurosci.*, 11:127–146.

Takayama, L. and Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5495 –5502.

Thomas, N. J. (2013). Mental imagery. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? an active perception approach to conscious mental content. *Cognitive Science*, 23:207–245.

Tomasello, M. (1995). Joint attention as social cognition. In Moore, C. and Dunham, P., editors, *Joint attention: its origins and role in development*, pages 103–130. Lawrence Erlbaum Associates.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–691.

Tourville, J. A., Reilly, K. J., and Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, pages 1429–1443.

Treisman, A. M. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177.

Tucker, M. and Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of experimental psychology. Human perception and performance*, 24:830–846.

Vallejo, G., Fernández, M. P., Tuero, E., and Livacic-Rojas, P. (2010). Análisis de medidas repetidas usando métodos de remuestreo (analyzing repeated measures using resampling methods). *Anales de Psicologia*, 26(2):400–409.

van Oosterhout, T. and Visser, A. (2008). A visual method for robot proxemics measurements. *Proceedings of the Metrics for Human-Robot Interaction Workshop in affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI 2008), Technical Report 471*, pages 61–68.

Varela, F. J., Thompson, E. T., and Rosch, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, new edition edition.

von Hofsten, C. (1982). Eyehand coordination in the newborn. *Developmental Psychology*, 18:450–461.

Walton, G. E., Bower, N., and Bower, T. (1992). Recognition of familiar faces by newborns. *Infant Behavior and Development*, 15(2):265 – 269.

Watson, J. B. (1913). Psychology as the behaviorist sees it. *Psychological Review*, 20:158–177.

Wells, G. L. and Petty, R. E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1:219–230.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291(5504):599–600.

Wexler, M. and Klam, F. (2001). Movement prediction and movement production. *Journal of Experimental Psychology: Human Perception and Performance*, 27:48–64.

Wexler, M., Kosslyn, S. M., and Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68(1):77 – 94.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9:625–636.

Wilson, M. and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131:460–473.

Wilson, R. A. and Foglia, L. (2011). Embodied cognition. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Winograd, T. and Flores, F. (1987). *Understanding computers and cognition: a new foundation for design*. Addison-Wesley.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Phil. Trans. of the Royal Soc. of London. Series B: Biol. Sc.*, 358(1431):593–602.

Wolpert, D. M. and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3 Suppl:1212–1217.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269:1880.

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7–8):1317 – 1329.

Ziemke, T., Jirenhed, D.-A., and Hesslow, G. (2005). Internal simulation of perception: a minimal neuro-robotic model. *Neurocomput.*, 68:85–104.

Zoia, S., Blason, L., DOttavio, G., Bulgheroni, M., Pezzetta, E., Scabar, A., and Castiello, U. (2007). Evidence of early development of action planning in the human foetus: a kinematic study. *Experimental Brain Research*, 176(2):217–226.

Zwickel, J., Hegele, M., and Grosjean, M. (2012). Ocular tracking of biological and non-biological motion: The effect of instructed agency. *Psychonomic Bulletin and Review*, 19(1):52–57.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation mit dem Titel "Sensorimotor Learning and Simulation of Experience as a Basis for the Development of Cognition in Robotics" selbstständig und ohne unerlaubte Hilfe angefertigt habe.
Ich habe die Arbeit nicht bereits an einer anderen Universität eingereicht und besitze keinen Doktorgrad im Fach Informatik.
Die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II vom 17.1. 2005, zuletzt geändert am 13.6.2006, veröffentlicht im Amtlichen Mitteilungsblatt der HU Nr. 34/2006, ist mir bekannt.

Berlin, den

Guido Schillaci